Representer Theorem for Learning Koopman Operators

Khosravi, Mohammad

DOI
10.1109/TAC.2023.3242325

Publication date
2023

Document Version
Final published version

Published in
IEEE Transactions on Automatic Control

Representer Theorem for Learning Koopman Operators

Mohammad Khosravi, *Member, IEEE*

***Abstract*—In this work, we consider the problem of learning the Koopman operator for discrete-time autonomous systems. The learning problem is formulated as a generic constrained regularized empirical loss minimization in the infinite-dimensional space of linear operators. We show that a representer theorem holds for the introduced learning problem under certain but general conditions, which allows convex reformulation of the problem in a specific finite-dimensional space without any approximation and loss of precision. We discuss the inclusion of various forms of regularization and constraints in the learning problem, such as the operator norm, the Frobenius norm, the operator rank, the nuclear norm, and the stability. Subsequently, we derive the corresponding equivalent finite-dimensional problem. Furthermore, we demonstrate the connection between the proposed formulation and the extended dynamic mode decomposition. We present several numerical examples to illustrate the theoretical results and verify the performance of regularized learning of the Koopman operators.**

***Index Terms*—Koopman operators, learning, representer theorem.**

I. Introduction

LEARNING-BASED and data-driven approaches for analysis, modeling, and control of nonlinear dynamics have received considerable attention in the recent years [1], [2], [3], [4]. In these approaches, various tools are developed to systematically analyze nonlinear dynamical systems based on the collected data and exploit valuable information and features such as stability. The main root of these techniques is in the classical point of view of the dynamical systems in which the system is described based on its state-space model. Considering that the evolution in nonlinear systems is defined through nonlinear maps characterizing the difference or differential equation of the system, learning methods are employed for obtaining these governing rules from the available measurement or synthetic data. These methodologies include regression techniques in the reproducing kernel Hilbert spaces (RKHS), polynomial optimization methods, Gaussian process regression, training neural networks, and many other ones [5], [6], [7], [8], [9], [10]. Working directly with the time-series data generated by the system has the leverage of dealing with finite-dimensional objects; however, the nonlinearity of the dynamics and incomplete information about the state trajectories can be challenging issues.

In addition to the approaches mentioned above, recent efforts in learning-based analysis of nonlinear dynamics have focused on the data-driven operator-theoretic methods, which lift the dynamics to an infinite-dimensional space via linear embeddings. More precisely, along with each dynamical system, a so-called *Koopman operator* is introduced that translates the dynamics of the system to a linear dynamical system on an infinite-dimensional space of functions called observables [11], [12]. While this framework leads to working in an infinite-dimensional space, we only need to deal with linear maps, and thus, one can leverage the linearity of resulting objects and employ well-developed tools of functional analysis and operator theory. The main feature of this lifting is the potential for full representation and global characterization of the underlying dynamical system [12], [13], [14]. This alternative formalism offers applicability in data-driven settings for the analysis of large classes of nonlinear and high-dimensional systems [11], [15], e.g., the framework has been applied to systems with different features, such as hyperbolic fixed points, limit cycles, and, attractors [16]. Besides the analysis of nonlinear dynamical systems, Koopman linear representation is also used for control, whereby not only the problem simplifies but also, in certain cases, it can outperform feedback policies designed based on the underlying nonlinear dynamics [17], [18], [19]. In data-driven settings, Koopman operators are already utilized in many applications, such as robotics [20], [21], human locomotion [22], neuroscience [23], fluid mechanics [24], and climate forecast [25]. Moreover, inspired by the physics of these applications or the expected behavior of the underlying system, different features, constraints, and forms of a priori information about the original system have been included in learning Koopman operators [26], [27], [28], [29]. For example, the stability constraints are imposed on the data-driven approximations of Koopman operators in [28] to address the long-term accuracy issue. In [29], the dissipativity constraint, which is a physics-based property of the system, is imposed in the approximation of the corresponding Koopman operator.

While the mentioned operator-theoretic approach has various potentials and benefits, its underlying infinite-dimensional

Manuscript received 3 August 2022; revised 13 January 2023; accepted 17 January 2023. Date of publication 6 February 2023; date of current version 26 April 2023. Recommended by Senior Editor Tetsuya Iwasaki and Guest Editors George J. Pappas, Anuradha M. Annaswamy, Manfred Morari, Claire J. Tomlin, Rene Vidal, and Melanie N. Zeilinger.

The author is with the Delft Center for Systems and Control, Delft University of Technology, 2628 CD Delft, The Netherlands (e-mail: mohammad.khosravi@tudelft.nl).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TAC.2023.3242325.

Digital Object Identifier 10.1109/TAC.2023.3242325

nature hinders its practical applicability unless one can obtain a suitable finite-dimensional approximation for the Koopman operator [17], [27], [30], [31]. To address this issue, various methods are developed, including dynamic mode decomposition (DMD) [32], Hankel-DMD [33], extended DMD (EDMD) [26], [34], generator EDMD (gEDMD) [35], to name a few. These linearization methods are either locally accurate or depend strongly on the choice of observable functions to provide suitable accuracy. Moreover, in all of the abovementioned approaches, rather than direct learning the infinite-dimensional Koopman operator, they initially enforce a restriction to a finite-dimensional subspace, and then, the data are employed to learn the corresponding finite-dimensional matrix representation. Accordingly, they suffer from a systematic loss of precision and inefficient utilization and exploitation of data. On the other hand, problem formulations for learning infinite-dimensional objects are generally ill-posed and intractable. Similar concerns arise in the most existing nonparametric learning problems. In those contexts, the issue is addressed using well-known representer theorems [36], [37], [38], [39], which provide conditions for characterizing the solution of learning problems as a finite linear combination of known objects. This highlights the necessity of developing analogous mathematical results for learning infinite-dimensional Koopman operators.

Motivated by the above discussion, we propose an operator-theoretic approach for learning the Koopman operator, formulated directly in the infinite-dimensional space of linear operators. The learning problem, in its most general form, is introduced as a constrained regularized empirical loss minimization, where the constraints enforce attributes of interests and potentially available side-information about the unknown Koopman operator, and the regularization is considered to penalize undesired features and avoid overfitting. We address the main concerns of data efficiency and tractability by introducing a set of representer theorems that provide equivalent finite-dimensional optimization problems. We first consider the case where the operator norm is employed for the regularization. We extend the results to the formulation in which linear constraints are additionally imposed enforcing features of interest. Subsequently, we demonstrate the connection between the proposed Koopman learning formulation and the well-known EDMD method. Then, we generalize the developed representer theorem and provide conditions under which the results hold. Following this, we consider different cases of regularization and constraints, e.g., Frobenius norm, nuclear norm, rank, and stability. For each of the resulting Koopman learning problems, we derive the equivalent finite-dimensional version that can be solved in a computationally tractable way using standard convex optimization techniques and singular value decomposition (SVD). Finally, we provide several illustrative numerical examples.

II. Notation and Preliminaries

The set of natural numbers, the set of nonnegative integers, the set of real scalars, the set of nonnegative real numbers, the n-dimensional Euclidean space, and the space of n by m real matrices are denoted by $\mathbb{N}$, $\mathbb{Z}_+$, $\mathbb{R}$, $\mathbb{R}_+$, $\mathbb{R}^n$, and $\mathbb{R}^{n\times m}$, respectively. For each $n \in \mathbb{N}$, $\{1, \ldots, n\}$ is denoted by $[n]$. For

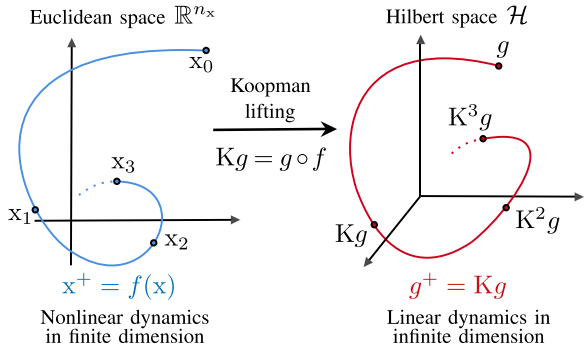

Fig. 1. Lifting of the dynamics from finite-dimensional state space to the infinite-dimensional Hilbert space of observables.

matrix $\mathrm{A} \in \mathbb{R}^{n\times m}$, the entry at the ith row and the jth column is denoted by $[\mathrm{A}]_{(i,j)}$. Given Hilbert spaces $\mathcal{H}$ and $\mathcal{W}$, the space of bounded linear operators $\mathrm{T} : \mathcal{H} \to \mathcal{W}$ is denoted by $\mathcal{L}(\mathcal{H}, \mathcal{W})$, and when $\mathcal{W}$ is the same as $\mathcal{H}$, we simply use $\mathcal{L}(\mathcal{H})$. Given vectors $u, v \in \mathcal{H}$, one can define a rank-one bounded linear operator, denoted by $v \otimes u$, such that $(v \otimes u)w = v\langle u, w\rangle$, for any $w \in \mathcal{H}$. For vectors $v_1, \ldots, v_n \in \mathcal{H}$, the Gram matrix $\mathrm{V} \in \mathbb{R}^{n\times n}$ is defined as $[\mathrm{V}]_{(i,j)} = \langle v_i, v_j\rangle$, for $i, j = 1, \ldots, n$. For matrix or operator A, the Frobenius or the Hilbert–Schmidt norm, the trace, the nuclear norm, and the adjoint are, respectively, denoted by $\|\mathrm{A}\|_{\mathrm{F}}$, $\mathrm{tr}(\mathrm{A})$, $\|\mathrm{A}\|_*$, and A^*. We have $(u \otimes v)^* = v \otimes u$. Let $\mathcal{H}$ contain $\mathbb{V}$-valued functions defined on $\mathcal{X}$, where $\mathcal{X}$ is a given set and $\mathbb{V}$ is a normed space. Then, for each $\mathrm{x} \in \mathcal{X}$, the *evaluation* operator at x, denoted by e_x, is a linear map $\mathrm{e}_\mathrm{x} : \mathcal{H} \to \mathbb{V}$ such that $\mathrm{e}_\mathrm{x}(g) = g(\mathrm{x})$, for any $g \in \mathcal{H}$. The RKHS with kernel $\Bbbk : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is denoted by $\mathcal{H}_\Bbbk$. Given $\mathrm{x} \in \mathcal{X}$, the *section of kernel* at x is defined as function $\Bbbk(\mathrm{x}, \cdot) : \mathcal{X} \to \mathbb{R}$. The domain of function $f : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$, denoted by $\mathrm{dom}(f)$, is defined as $\mathrm{dom}(f) := \{\mathrm{x} \in \mathcal{X} \,|\, f(\mathrm{x}) < \infty\}$. Given sets $\mathcal{B}$ and $\mathcal{C}$ such that $\mathcal{C} \subset \mathcal{B}$, the indicator function $\delta_\mathcal{C} : \mathcal{B} \to \mathbb{R} \cup \{+\infty\}$ is defined as $\delta_\mathcal{C}(a) = 0$, when $a \in \mathcal{C}$, and $\delta_\mathcal{C}(a) = +\infty$, if $a \notin \mathcal{C}$.

III. Dynamical Systems and Koopman Operators

Let $f : \mathcal{X} \to \mathcal{X}$ be a map defined over $\mathcal{X} \subseteq \mathbb{R}^{n_\mathrm{x}}$. Consider the following autonomous dynamical system:

$$\mathrm{x}_{k+1} = f(\mathrm{x}_k), \qquad k \in \mathbb{Z}_+ \tag{1}$$

characterizing a discrete-time flow on the state space $\mathcal{X}$. Let $\mathcal{H}$ be a Hilbert space of real-valued functions defined over $\mathcal{X}$. Suppose $\mathcal{H}$ is closed under composition with map f, i.e., we know that $g \circ f \in \mathcal{H}$, for each $\mathrm{g} \in \mathcal{H}$. Given a trajectory of the system, with respect to each $g \in \mathcal{H}$, we have a sequence of real numbers transferring information about system (1) through the lens of g. Accordingly, each element of $\mathcal{H}$ is called an *observable*.

Definition 1 (Koopman operator): With respect to dynamical system (1), the *Koopman* operator is defined as a linear map $\mathrm{K} : \mathcal{H} \to \mathcal{H}$ such that $\mathrm{K}g = g \circ f$, for each $g \in \mathcal{H}$.

The Koopman operator induces a lifting of the finite-dimensional nonlinear dynamics (1) to the infinite-dimensional linear dynamics $g^+ = \mathrm{K}g$, which is in the Hilbert space of observables $\mathcal{H}$ (see Fig. 1). Accordingly, the dynamical system (1) can be completely described based on the corresponding

Koopman operator K [11]. Motivated by this fact, when a set of data about the dynamical system (1) is provided through several observables, we can pose the problem of learning the Koopman operator using the given data and potentially available side-information. To this end, we propose an operator-theoretic framework in Section IV, formulating the Koopman operator learning problem as a generic infinite-dimensional constrained regularized empirical loss minimization. Theorem 1 in Section V provides a representer theorem for the case of Tikhonov regularization, which also shows that the learning problem is well-posed and tractable. In Section VI, Theorem 3 extends this result to the case where we have additional linear constraints, and following this, the connection to the EDMD method is elaborated in Section VII. The most general case is studied by Theorem 8, and subsequently, various cases of interest are discussed in Section VIII.

IV. Learning Koopman Operator

Let $\mathrm{x}_0, \mathrm{x}_1, \ldots, \mathrm{x}_{n_\mathrm{s}}$ be a trajectory of dynamical system (1) and $g_1, g_2, \ldots, g_{n_\mathrm{g}} \in \mathcal{H}$ be a set of observable maps. Accordingly, set of data $\mathscr{D}$ is provided as

$$\mathscr{D} := \left\{ y_{kl} := g_l(\mathrm{x}_k) \,\middle|\, k = 0, \ldots, n_\mathrm{s}, l = 1, \ldots, n_\mathrm{g} \right\}. \tag{2}$$

Note that one may define y_{kl} as $y_{kl} = g_l(\mathrm{x}_k) + \varepsilon_{kl}$, for $k = 0, \ldots, n_\mathrm{s}$ and $l = 1, \ldots, n_\mathrm{g}$, where ε_{kl} is introduced for considering the possible uncertainties in the value of observable g_l at x_k that can be potentially due to imperfect measurements or evaluations. We can also apply similar considerations to the trajectory data. In order to learn the Koopman operator of dynamical system (1), we need a suitable learning objective function to be minimized over the hypothesis space of candidate operators $\mathcal{L}(\mathcal{H})$. To this end, we define *regularized empirical loss* function, $\mathcal{J} : \mathcal{L}(\mathcal{H}) \to \mathbb{R}_+ \cup \{+\infty\}$, as

$$\mathcal{J}(\mathrm{K}) := \mathcal{E}(\mathrm{K}) + \lambda \mathcal{R}(\mathrm{K}) \tag{3}$$

where $\mathcal{R} : \mathcal{L}(\mathcal{H}) \to \mathbb{R}_+ \cup \{+\infty\}$ is a regularization function, $\lambda > 0$ is the weight of regularization, and $\mathcal{E} : \mathcal{L}(\mathcal{H}) \to \mathbb{R}_+$ is the *empirical loss* function characterized as

$$\mathcal{E}(\mathrm{K}) := \sum_{k=1}^{n_\mathrm{s}} \sum_{l=1}^{n_\mathrm{g}} \left(y_{kl} - (\mathrm{K} g_l)(\mathrm{x}_{k-1}) \right)^2. \tag{4}$$

Furthermore, we may also consider a constraint set $\mathcal{C} \subseteq \mathcal{L}(\mathcal{H})$ in the learning problem for the inclusion of some possibly available side-information on the Koopman operator. This side-information can be about different aspects of the local or global behavior of the dynamical system (1), e.g., the stability of an equilibrium point. Accordingly, the optimization problem for learning the Koopman operator is defined as

$$\begin{aligned} \min_{\mathrm{K} \in \mathcal{L}(\mathcal{H})} \quad & \mathcal{J}(\mathrm{K}) = \mathcal{E}(\mathrm{K}) + \lambda \mathcal{R}(\mathrm{K}) \\ \text{s.t.} \quad & \mathrm{K} \in \mathcal{C}. \end{aligned} \tag{5}$$

Note that (5) is an optimization problem over an infinite-dimensional set. Hence, the main concern is whether this problem admits a solution, and if such solution exists, how one can obtain it in a computationally tractable way. These issues depend on the choice of the regularization function and the constraint set. In the following, we study various settings and provide conditions under which a representer theorem holds for the learning problem (3) and also for its generalized form.

V. Learning Koopman Operator With Tikhonov Regularization

In the statistical learning theory, Tikhonov regularization is the most common choice, i.e., $\mathcal{R}$ is defined as the quadratic function $\mathcal{R}(\mathrm{K}) := \|\mathrm{K}\|^2$. If there is no additional constraint on the Koopman operator, we have the following learning problem:

$$\min_{\mathrm{K} \in \mathcal{L}(\mathcal{H})} \sum_{k=1}^{n_\mathrm{s}} \sum_{l=1}^{n_\mathrm{g}} \left(y_{kl} - (\mathrm{K} g_l)(\mathrm{x}_{k-1}) \right)^2 + \lambda \|\mathrm{K}\|^2 \tag{6}$$

which is a special case of (5). To study this problem, we need a technical assumption given below.

Assumption 1: For $k = 0, \ldots, n_\mathrm{s} - 1$, the evaluation operator $\mathrm{e}_{\mathrm{x}_k}$ is continuous (bounded), i.e., $\mathrm{e}_{\mathrm{x}_k} \in \mathcal{L}(\mathcal{H}, \mathbb{R})$.

This assumption says that, for each $k \in \{0, \ldots, n_\mathrm{s} - 1\}$, the value of $g(\mathrm{x}_k)$ depends continuously on g. More precisely, there exists a nonnegative constant C such that we have $|g(\mathrm{x}_k)| \leq C\|g\|$, for each $g \in \mathcal{H}$. Roughly speaking, by a small perturbation of g, we expect that the value of $g(\mathrm{x}_k)$ does not change significantly. Namely, when $g, h \in \mathcal{H}$ are almost similar observables, i.e., $\|g - h\|$ is small, $g(\mathrm{x}_k)$ and $h(\mathrm{x}_k)$ are expected to have almost similar values. Accordingly, one can see that Assumption 1 is quite natural, otherwise, the observable functions in $\mathcal{H}$ do not provide reliable and useful information on dynamics (1).

Remark 1: A special case where Assumption 1 holds is when $\mathcal{H}$ is an RKHS [40], [41]. More precisely, if $\mathcal{H}$ is endowed with *reproducing kernel* $\Bbbk : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, we know that

$$|g(\mathrm{x}_k)| = |\langle \Bbbk(\mathrm{x}_k, \cdot), g \rangle| \leq \|\Bbbk(\mathrm{x}_k, \cdot)\| \|g\|, \qquad \forall g \in \mathcal{H} \tag{7}$$

which is due to the *reproducing property* of the kernel and the Cauchy–Schwartz inequality. Accordingly, if C is defined as

$$C := \max \left\{ \|\Bbbk(\mathrm{x}_k, \cdot)\| \,\middle|\, k = 0, \ldots, n_\mathrm{s} - 1 \right\} \tag{8}$$

then, for any $g \in \mathcal{H}$, we have $|g(\mathrm{x}_k)| \leq C\|g\|$.

The next theorem characterizes the solution of (6). More precisely, we derive an equivalent finite-dimensional convex program to solve infinite-dimensional optimization problem (6) without any approximation error.

Theorem 1: Let Assumption 1 hold and $\lambda > 0$. Then, the optimization problem (6) admits a unique solution denoted by $\hat{\mathrm{K}}$. Moreover, there exist vectors $v_1, \ldots, v_{n_\mathrm{s}} \in \mathcal{H}$ such that

$$\hat{\mathrm{K}} = \sum_{k=1}^{n_\mathrm{s}} \sum_{l=1}^{n_\mathrm{g}} a_{kl}\, v_k \otimes g_l \tag{9}$$

where $\mathrm{A} = [a_{kl}]_{k=1,l=1}^{n_\mathrm{s}, n_\mathrm{g}} \in \mathbb{R}^{n_\mathrm{s} \times n_\mathrm{g}}$ is the solution of the following optimization problem:

$$\min_{\mathrm{A} \in \mathbb{R}^{n_\mathrm{s} \times n_\mathrm{g}}} \|\mathrm{VAG} - \mathrm{Y}\|_\mathrm{F}^2 + \lambda \|\mathrm{V}^{\frac{1}{2}} \mathrm{A} \mathrm{G}^{\frac{1}{2}}\|^2 \tag{10}$$

given that V and G are, respectively, the Gramian matrix of $\{v_1, \ldots, v_{n_\mathrm{s}}\}$ and $\{g_1, \ldots, g_{n_\mathrm{g}}\}$, and $\mathrm{Y} \in \mathbb{R}^{n_\mathrm{s} \times n_\mathrm{g}}$ is the matrix defined as $\mathrm{Y} := [y_{kl}]_{k=1,l=1}^{n_\mathrm{s}, n_\mathrm{g}}$.

Proof: Since operators $\mathrm{e}_{\mathrm{x}_0}, \ldots, \mathrm{e}_{\mathrm{x}_{n_\mathrm{s}-1}}$ are bounded, due to the Riesz representation theorem [42], we know that there exists $v_k \in \mathcal{H}$ such that $\mathrm{e}_{\mathrm{x}_{k-1}}(\cdot) = \langle v_k, \cdot \rangle$, for $k \in [n_\mathrm{s}]$. Therefore, we

have $(\mathrm{K}g_l)(\mathrm{x}_{k-1}) = \langle v_k, \mathrm{K}g_l\rangle$, for each $k \in [n_\mathrm{s}]$ and each $l \in [n_\mathrm{g}]$. Thus, one can write the objective function in (6) as

$$\mathcal{J}(\mathrm{K}) := \sum_{k=1}^{n_\mathrm{s}}\sum_{l=1}^{n_\mathrm{g}}(y_{kl} - \langle v_k, \mathrm{K}g_l\rangle)^2 + \lambda\|\mathrm{K}\|^2. \tag{11}$$

Since $\langle v_k, \mathrm{K}g_l\rangle$ is linear and continuous with respect to K, we have that $\mathcal{J} : \mathcal{L}(\mathcal{H}) \to \mathbb{R}$ is a continuous map. Also, as $\lambda > 0$, one can see that $\mathcal{J}$ is a strongly convex function. Therefore, the optimization problem $\min_{\mathrm{K}\in\mathcal{L}(\mathcal{H})} \mathcal{J}(\mathrm{K})$ admits a unique solution [43] denoted by $\hat{\mathrm{K}}$.

Let the linear subspaces $\mathcal{V}$ and $\mathcal{G}$ be defined, respectively, as $\mathcal{V} := \mathrm{span}\{v_1, \ldots, v_{n_\mathrm{s}}\}$ and $\mathcal{G} := \mathrm{span}\{g_1, \ldots, g_{n_\mathrm{g}}\}$. Also, let $\Pi_\mathcal{V} : \mathcal{H} \to \mathcal{H}$ and $\Pi_\mathcal{G} : \mathcal{H} \to \mathcal{H}$ denote the projection operator on $\mathcal{V}$ and $\mathcal{G}$, respectively. Since the dimension of $\mathcal{V}$ and $\mathcal{G}$ are finite, they are closed subspaces of $\mathcal{H}$, and hence, the projection operators $\Pi_\mathcal{V}$ and $\Pi_\mathcal{G}$ are well-defined. Let operator S be defined as $\mathrm{S} := \Pi_\mathcal{V}\hat{\mathrm{K}}\Pi_\mathcal{G}$. Due to the definition of $\Pi_\mathcal{G}$, we know that $\Pi_\mathcal{G} g_l = g_l$, for $l \in [n_\mathrm{g}]$. Accordingly, for each k and l, we have

$$\begin{aligned}\langle v_k, \mathrm{S}g_l\rangle &= \left\langle v_k, \Pi_\mathcal{V}\hat{\mathrm{K}}\Pi_\mathcal{G} g_l\right\rangle \\ &= \left\langle v_k, \Pi_\mathcal{V}\hat{\mathrm{K}}g_l\right\rangle = \left\langle \Pi_\mathcal{V}^* v_k, \hat{\mathrm{K}}g_l\right\rangle\end{aligned} \tag{12}$$

where $\Pi_\mathcal{V}^*$ is the adjoint of $\Pi_\mathcal{V}$. Since $\mathcal{V}$ is a closed subspace, the projection operator $\Pi_\mathcal{V}$ is self-adjoint, i.e., $\Pi_\mathcal{V}^* = \Pi_\mathcal{V}$. Hence, for each $k \in [n_\mathrm{s}]$, we have $\Pi_\mathcal{V}^* v_k = \Pi_\mathcal{V} v_k = v_k$, where the second equality is due to the definition of operator $\Pi_\mathcal{V}$. Accordingly, from (12), we know that $\langle v_k, \mathrm{S}g_l\rangle = \langle \Pi_\mathcal{V} v_k, \hat{\mathrm{K}}g_l\rangle = \langle v_k, \hat{\mathrm{K}}g_l\rangle$, for each $k \in [n_\mathrm{s}]$ and $l \in [n_\mathrm{g}]$. Consequently, it follows that

$$\sum_{k=1}^{n_\mathrm{s}}\sum_{l=1}^{n_\mathrm{g}}(y_{kl} - \langle v_k, \mathrm{S}g_l\rangle)^2 = \sum_{k=1}^{n_\mathrm{s}}\sum_{l=1}^{n_\mathrm{g}}\left(y_{kl} - \langle v_k, \hat{\mathrm{K}}g_l\rangle\right)^2. \tag{13}$$

Since $\Pi_\mathcal{V}$ and $\Pi_\mathcal{G}$ are projection operators, we know that $\|\Pi_\mathcal{V}\|, \|\Pi_\mathcal{G}\| \le 1$. Therefore, we have $\|\mathrm{S}\|^2 \le \|\hat{\mathrm{K}}\|^2$. Consequently, from (13), one can see that $\mathcal{J}(\mathrm{S}) \le \mathcal{J}(\hat{\mathrm{K}})$. Since, the operator $\hat{\mathrm{K}}$ is the unique solution of $\min_{\mathrm{K}\in\mathcal{L}(\mathcal{H})} \mathcal{J}(\mathrm{K})$, we need to have $\hat{\mathrm{K}} = \mathrm{S} = \Pi_\mathcal{V}\hat{\mathrm{K}}\Pi_\mathcal{G}$. Due to the linearity of operator $\hat{\mathrm{K}}$, it follows that there exist $a_{kl} \in \mathbb{R}$, for $k \in [n_\mathrm{s}]$ and $l \in [n_\mathrm{g}]$, such that $\hat{\mathrm{K}} = \sum_{k=1}^{n_\mathrm{s}}\sum_{l=1}^{n_\mathrm{g}} a_{kl} v_k \otimes g_l$. To find these values, we replace K in $\min_{\mathrm{K}\in\mathcal{L}(\mathcal{H})} \mathcal{J}(\mathrm{K})$ with $\hat{\mathrm{K}}$ considering the given parametric form. To this end, we need to calculate the value of empirical loss and the regularization term for $\hat{\mathrm{K}}$. Note that, for each $k \in [n_\mathrm{s}]$ and $l, j \in [n_\mathrm{g}]$, we have $(v_k \otimes g_l)g_j = v_k\langle g_l, g_j\rangle$. Accordingly, due to the linearity of inner product, for $\hat{\mathrm{K}} = \sum_{k=1}^{n_\mathrm{s}}\sum_{l=1}^{n_\mathrm{g}} a_{kl}\, v_k \otimes g_l$, we have

$$\begin{aligned}(\hat{\mathrm{K}}g_j)(\mathrm{x}_{i-1}) &= \sum_{k=1}^{n_\mathrm{s}}\sum_{l=1}^{n_\mathrm{g}} a_{kl}\,\langle v_i, (v_k \otimes g_l)g_j\rangle \\ &= \sum_{k=1}^{n_\mathrm{s}}\sum_{l=1}^{n_\mathrm{g}} a_{kl}\,\langle v_i, v_k\langle g_l, g_j\rangle\rangle = \sum_{k=1}^{n_\mathrm{s}}\sum_{l=1}^{n_\mathrm{g}}\langle v_i, v_k\rangle\, a_{kl}\,\langle g_l, g_j\rangle\end{aligned}$$

for each $i \in [n_\mathrm{s}]$ and $j \in [n_\mathrm{g}]$. Subsequently, due to the definition of Gramian matrices V and G, it follows that

$$\left(\hat{\mathrm{K}}g_j\right)(\mathrm{x}_{i-1}) = [\mathrm{VAG}]_{(i,j)}. \tag{14}$$

Therefore, given the definition of matrix Y, we have

$$\begin{aligned}&\textstyle\sum_{k=1}^{n_\mathrm{s}}\sum_{l=1}^{n_\mathrm{g}}\left(y_{kl} - (\hat{\mathrm{K}}g_l)(\mathrm{x}_{k-1})\right)^2 \\ &= \textstyle\sum_{k=1}^{n_\mathrm{s}}\sum_{l=1}^{n_\mathrm{g}}\left([\mathrm{Y}]_{(k,l)} - [\mathrm{VAG}]_{(k,l)}\right)^2 = \|\mathrm{Y} - \mathrm{VAG}\|_\mathrm{F}^2\end{aligned} \tag{15}$$

where A is the matrix defined as $\mathrm{A} = [a_{kl}]_{k=1,l=1}^{n_\mathrm{s},\,n_\mathrm{g}}$. We also need to derive the value of $\|\hat{\mathrm{K}}\|^2$. For each $h \in \mathcal{H}$, we have

$$\hat{\mathrm{K}}h = \left(\sum_{k=1}^{n_\mathrm{s}}\sum_{l=1}^{n_\mathrm{g}} a_{kl}\, v_k \otimes g_l\right)h = \sum_{k=1}^{n_\mathrm{s}}\sum_{l=1}^{n_\mathrm{g}} a_{kl}\, v_k\langle g_l, h\rangle.$$

We know that $\langle g_l, h\rangle = \langle g_l, \Pi_\mathcal{G} h\rangle$, for each $l \in [n_\mathrm{g}]$. Accordingly, since $\Pi_\mathcal{G} h \in \mathcal{G} = \mathrm{span}\{g_l\}_{l=1}^{n_\mathrm{g}}$ and due to the definition of operator norm, one has

$$\begin{aligned}\|\hat{\mathrm{K}}\|^2 &= \sup_{h\in\mathcal{H}} \frac{\|\sum_{k=1}^{n_\mathrm{s}}\sum_{l=1}^{n_\mathrm{g}} a_{kl}\, v_k\langle g_l, \Pi_\mathcal{G} h\rangle\|^2}{\|\Pi_\mathcal{G} h\|^2 + \|\Pi_{\mathcal{G}^\perp} h\|^2} \\ &= \sup_{\mathrm{c}\in\mathbb{R}^{n_\mathrm{g}}} \frac{\|\sum_{k=1}^{n_\mathrm{s}}\sum_{l=1}^{n_\mathrm{g}} a_{kl}\, v_k\langle g_l, \sum_{j=1}^{n_\mathrm{g}} c_j g_j\rangle\|^2}{\|\sum_{j=1}^{n_\mathrm{g}} c_j g_j\|^2}\end{aligned}$$

where $\mathrm{c} = [c_1, \ldots, c_{n_\mathrm{g}}]^\mathsf{T} \in \mathbb{R}^{n_\mathrm{g}}$ is the vector of coefficients in the expansion of $\Pi_\mathcal{G} h$ as in $\Pi_\mathcal{G} h = \sum_{j=1}^{n_\mathrm{g}} c_j g_j$. Note that we have

$$\left\|\sum_{j=1}^{n_\mathrm{g}} c_j g_j\right\|^2 = \sum_{j_1=1}^{n_\mathrm{s}}\sum_{j_2=1}^{n_\mathrm{s}} c_{j_1}\langle g_{j_1}, g_{j_2}\rangle c_{j_2} = \mathrm{c}^\mathsf{T}\mathrm{G}\mathrm{c}.$$

Also, due to the linearity of inner product, one can see that

$$\begin{aligned}\sum_{k=1}^{n_\mathrm{s}}\sum_{l=1}^{n_\mathrm{g}} a_{kl}\, v_k\langle g_l, \sum_{j=1}^{n_\mathrm{g}} c_j g_j\rangle &= \sum_{k=1}^{n_\mathrm{s}}\sum_{l=1}^{n_\mathrm{g}}\sum_{j=1}^{n_\mathrm{g}} a_{kl} v_k c_j\langle g_l, g_j\rangle \\ &= \sum_{k=1}^{n_\mathrm{s}} v_k d_k\end{aligned}$$

where d_k is defined as $d_k := \sum_{l=1}^{n_\mathrm{g}}\sum_{j=1}^{n_\mathrm{g}} a_{kl}\langle g_l, g_j\rangle c_j$, for each $k \in [n_\mathrm{s}]$. Accordingly, we have

$$\left\|\sum_{k=1}^{n_\mathrm{s}}\sum_{l=1}^{n_\mathrm{g}} a_{kl}\, v_k\langle g_l, \sum_{j=1}^{n_\mathrm{g}} c_j g_j\rangle\right\|^2 = \mathrm{d}^\mathsf{T}\mathrm{V}\mathrm{d}$$

where V is the matrix defined as $\mathrm{V} = [\langle v_{k_1}, v_{k_2}\rangle]_{k_1,k_2=1}^{n_\mathrm{s},n_\mathrm{s}}$, and d is the vector defined as $\mathrm{d} := [d_k]_{k=1}^{n_\mathrm{s}}$. One can see that $\mathrm{d} = \mathrm{AG}\mathrm{c}$. Subsequently, we have

$$\begin{aligned}\|\hat{\mathrm{K}}\|^2 &= \sup_{\mathrm{c}\in\mathbb{R}^{n_\mathrm{g}}} \frac{\mathrm{c}^\mathsf{T}\mathrm{G}\mathrm{A}^\mathsf{T}\mathrm{V}\mathrm{A}\mathrm{G}\mathrm{c}}{\mathrm{c}^\mathsf{T}\mathrm{G}\mathrm{c}} \\ &= \sup_{\mathrm{c}\in\mathbb{R}^{n_\mathrm{g}}} \frac{\|\mathrm{V}^{\frac{1}{2}}\mathrm{A}\mathrm{G}^{\frac{1}{2}}\,\mathrm{G}^{\frac{1}{2}}\mathrm{c}\|^2}{\|\mathrm{G}^{\frac{1}{2}}\mathrm{c}\|^2} = \|\mathrm{V}^{\frac{1}{2}}\mathrm{A}\mathrm{G}^{\frac{1}{2}}\|^2\end{aligned} \tag{16}$$

where the last equality is implied from the definition of matrix norm and also using the change-of-variable $\mathrm{b} = \mathrm{G}^{\frac{1}{2}}\mathrm{c}$. Therefore, due to (15) and (16), one can see that

$$\mathcal{J}(\hat{\mathrm{K}}) = \|\mathrm{Y} - \mathrm{VAG}\|_\mathrm{F}^2 + \lambda\,\|\mathrm{V}^{\frac{1}{2}}\mathrm{A}\mathrm{G}^{\frac{1}{2}}\|^2. \tag{17}$$

This concludes the proof. ■

Remark 2: Let the Hilbert space of observables be the RKHS $\mathcal{H}_\Bbbk$. From the reproducing property of $\Bbbk$ [41], we have

$$\mathrm{e}_{\mathrm{x}_{k-1}}(g) = g(\mathrm{x}_{k-1}) = \langle \Bbbk(\mathrm{x}_{k-1}, \cdot), g\rangle \tag{18}$$

and subsequently, we have $\mathrm{v}_k = \Bbbk(\mathrm{x}_{k-1}, \cdot)$, for $k \in [n_\mathrm{s}]$. Following this, we can derive the Gramian matrix V as

$$\begin{aligned}[\mathrm{V}]_{(i,j)} &= \langle \mathrm{v}_i, \mathrm{v}_j\rangle \\ &= \langle \Bbbk(\mathrm{x}_{i-1}, \cdot), \Bbbk(\mathrm{x}_{j-1}, \cdot)\rangle = \Bbbk(\mathrm{x}_{i-1}, \mathrm{x}_{j-1})\end{aligned} \tag{19}$$

for each $i, j \in [n_{\mathrm{s}}]$, where the last equality is due to the reproducing property. Since the linear span of $\{\mathbb{k}_{\mathrm{x}} | \mathrm{x} \in \mathcal{X}\}$ is dense in $\mathcal{H}$, one may take observables as the sections of kernel at points of a given set $\mathcal{P} := \{\mathrm{p}_1, \ldots, \mathrm{p}_{n_{\mathrm{g}}}\} \subset \mathcal{X}$, i.e., $g_l(\cdot) := \mathbb{k}(\mathrm{p}_l, \cdot)$, for $l \in [n_{\mathrm{g}}]$. Then, similar to (19), one can see that $[\mathrm{G}]_{(l,j)} = \mathbb{k}(\mathrm{p}_l, \mathrm{p}_j)$, for any $l, j \in [n_{\mathrm{g}}]$. Moreover, we have $g_l(\mathrm{x}_{k-1}) = \mathbb{k}(\mathrm{p}_l, \mathrm{x}_{k-1})$, for any $k \in [n_{\mathrm{s}}+1]$ and $l \in [n_{\mathrm{g}}]$. One can also obtain similar expressions when observables are finite linear combinations of the sections of kernel.

Remark 3: The results introduced in Theorem 1 and the equivalent finite-dimensional optimization problem can be extended to the case where we have multiple trajectories of the system. More details are provided in [44, Appendix A].

VI. Learning Koopman Operator With Image in a Subspace of Interest

Let $\mathcal{W}$ be a closed linear subspace of $\mathcal{H}$, and define $\mathcal{L}_{\mathcal{W}}$ as the set of bounded linear operators mapping $\mathcal{G} = \operatorname{span}\{g_l\}_{l=1}^{n_{\mathrm{g}}}$ into $\mathcal{W}$, i.e.,

$$\mathcal{L}_{\mathcal{W}} := \{\mathrm{S} \in \mathcal{L}(\mathcal{H}) \mid \mathrm{S}(\mathcal{G}) \subseteq \mathcal{W}\}. \tag{20}$$

Since we may encode some specific form of prior knowledge through employing $\mathcal{W}$, it might be of particular interest to learn the Koopman operator as an element of $\mathcal{L}_{\mathcal{W}}$. For example, when $g_1, \ldots, g_{n_{\mathrm{g}}}$ and f are polynomials, respectively, with maximum degree of d_g and d_f, we know that $\mathrm{K}g_l = g_l \circ f$ is a polynomial with degree maximally equal to $d_g d_f$, for each $l = 1, \ldots, n_{\mathrm{g}}$. Accordingly, one may introduce $\mathcal{W}$ as the set of polynomials with degree less than or equal to $d_g d_f$. The next theorem provides the closest approximation of learned operator $\hat{\mathrm{K}}$, introduced in (9), in the closed subspace $\mathcal{L}_{\mathcal{W}}$.

Theorem 2: Define operator $\widetilde{\mathrm{K}}_{\mathcal{W}}$ as

$$\widetilde{\mathrm{K}}_{\mathcal{W}} := \sum_{k=1}^{n_{\mathrm{s}}} \sum_{l=1}^{n_{\mathrm{g}}} a_{kl} \left(\Pi_{\mathcal{W}} v_k\right) \otimes g_l \tag{21}$$

where $\mathrm{A} = [a_{kl}]_{k=1,l=1}^{n_{\mathrm{s}}, n_{\mathrm{g}}} \in \mathbb{R}^{n_{\mathrm{s}} \times n_{\mathrm{g}}}$ is the solution of (10) and $v_1, \ldots, v_{n_{\mathrm{s}}} \in \mathcal{H}$ are the vectors defined in Theorem 1. Then, we have $\widetilde{\mathrm{K}}_{\mathcal{W}} \in \mathcal{L}_{\mathcal{W}}$ and $\widetilde{\mathrm{K}}_{\mathcal{W}} \in \operatorname{argmin}_{\mathrm{S} \in \mathcal{L}_{\mathcal{W}}} \|\hat{\mathrm{K}} - \mathrm{S}\|$.

Proof: Being $\mathcal{W}$ a closed linear subspace of $\mathcal{H}$, the projection operator $\Pi_{\mathcal{W}}$ is well-defined. One can see that $\Pi_{\mathcal{W}}(v \otimes g) = (\Pi_{\mathcal{W}} v) \otimes g$, for any $v, g \in \mathcal{H}$. Accordingly, we have $\widetilde{\mathrm{K}}_{\mathcal{W}} = \Pi_{\mathcal{W}} \hat{\mathrm{K}}$, from which it follows that $\widetilde{\mathrm{K}}_{\mathcal{W}} \in \mathcal{L}_{\mathcal{W}}$. Since $\Pi_{\mathcal{G}}$ is a projection operator, we know that $\|\Pi_{\mathcal{G}}\| = 1$. Accordingly, for any $\mathrm{S} \in \mathcal{L}_{\mathcal{W}}$, we have

$$\|\hat{\mathrm{K}} - \mathrm{S}\Pi_{\mathcal{G}}\| = \|(\hat{\mathrm{K}} - \mathrm{S})\Pi_{\mathcal{G}}\| \leq \|\hat{\mathrm{K}} - \mathrm{S}\| \|\Pi_{\mathcal{G}}\| = \|\hat{\mathrm{K}} - \mathrm{S}\| \tag{22}$$

where the first equality is a result of $\hat{\mathrm{K}}\Pi_{\mathcal{G}} = \hat{\mathrm{K}}$, which is according to the definition of $\hat{\mathrm{K}}$ in (9). Also, due to the definition of $\mathcal{L}_{\mathcal{W}}$, we know that $\mathrm{S}(\mathcal{G}) \subseteq \mathcal{W}$, and consequently, one has $\Pi_{\mathcal{W}^\perp} \mathrm{S} \Pi_{\mathcal{G}} = 0$. Accordingly, it follows that

$$\begin{aligned} \|\Pi_{\mathcal{W}^\perp} \hat{\mathrm{K}}\| &= \|\Pi_{\mathcal{W}^\perp} \hat{\mathrm{K}} - \Pi_{\mathcal{W}^\perp} \mathrm{S} \Pi_{\mathcal{G}}\| \\ &\leq \|\Pi_{\mathcal{W}^\perp}\| \|\hat{\mathrm{K}} - \mathrm{S}\Pi_{\mathcal{G}}\| = \|\hat{\mathrm{K}} - \mathrm{S}\Pi_{\mathcal{G}}\| \end{aligned} \tag{23}$$

where the last equality is concluded from $\|\Pi_{\mathcal{W}^\perp}\| = 1$ that is due to the fact that $\Pi_{\mathcal{W}^\perp}$ is a projection operator. Furthermore, we know that $\Pi_{\mathcal{W}^\perp} = \mathbb{I} - \Pi_{\mathcal{W}}$, where $\mathbb{I}$ is the identity operator on $\mathcal{H}$. Hence, due to $\widetilde{\mathrm{K}}_{\mathcal{W}} = \Pi_{\mathcal{W}} \hat{\mathrm{K}}$, we have $\Pi_{\mathcal{W}^\perp} \hat{\mathrm{K}} = \hat{\mathrm{K}} - \widetilde{\mathrm{K}}_{\mathcal{W}}$. Therefore, from (22) and (23), it follows that

$$\|\hat{\mathrm{K}} - \widetilde{\mathrm{K}}_{\mathcal{W}}\| \leq \|\hat{\mathrm{K}} - \mathrm{S}\|, \qquad \forall \mathrm{S} \in \mathcal{L}_{\mathcal{W}} \tag{24}$$

which concludes the proof. ∎

According to Theorem 2, to obtain the closest approximation of operator $\hat{\mathrm{K}}$ in $\mathcal{L}_{\mathcal{W}}$, i.e., $\widetilde{\mathrm{K}}_{\mathcal{W}}$, we need to solve (10) and also derive $\Pi_{\mathcal{W}} v_k = \operatorname{argmin}_{w \in \mathcal{W}} \|v_k - w\|$, for $k = 1, \ldots, n_{\mathrm{s}}$. On the other hand, one may propose to learn the Koopman operator in $\mathcal{L}_{\mathcal{W}}$ via a direct approach by finding the solution of the following learning problem:

$$\min_{\mathrm{K} \in \mathcal{L}_{\mathcal{W}}} \ \mathcal{E}(\mathrm{K}) + \lambda \|\mathrm{K}\|^2 \tag{25}$$

where $\mathcal{E} : \mathcal{H} \to \mathbb{R}$ is the empirical loss defined in (4). The following theorem characterizes the solution of infinite-dimensional program (25) through obtaining an exact equivalent finite-dimensional convex reformulation.

Theorem 3: Let Assumption 1 hold, $\lambda > 0$, and $v_1, \ldots, v_{n_{\mathrm{s}}} \in \mathcal{H}$ be the vectors defined in Theorem 1. Then, the optimization problem (25) has a unique solution denoted by $\hat{\mathrm{K}}_{\mathcal{W}}$. Also, there exists $\mathrm{A} = [a_{kl}]_{k=1,l=1}^{n_{\mathrm{s}}, n_{\mathrm{g}}} \in \mathbb{R}^{n_{\mathrm{s}} \times n_{\mathrm{g}}}$, which characterizes $\hat{\mathrm{K}}_{\mathcal{W}}$ as

$$\hat{\mathrm{K}}_{\mathcal{W}} = \sum_{k=1}^{n_{\mathrm{s}}} \sum_{l=1}^{n_{\mathrm{g}}} a_{kl} \left(\Pi_{\mathcal{W}} v_k\right) \otimes g_l. \tag{26}$$

Define matrices G and Y as in Theorem 1, and $\mathrm{W}_{\mathcal{V}}$ as the Gramian matrix of vectors $\Pi_{\mathcal{W}} v_1, \ldots, \Pi_{\mathcal{W}} v_{n_{\mathrm{s}}}$. Then, matrix A in (26) is the solution of the following optimization problem:

$$\min_{\mathrm{A} \in \mathbb{R}^{n_{\mathrm{s}} \times n_{\mathrm{g}}}} \ \|\mathrm{W}_{\mathcal{V}} \mathrm{A}\mathrm{G} - \mathrm{Y}\|_{\mathrm{F}}^2 + \lambda \|\mathrm{W}_{\mathcal{V}}^{\frac{1}{2}} \mathrm{A}\mathrm{G}^{\frac{1}{2}}\|^2. \tag{27}$$

Proof: See Appendix A. ∎

In the next theorem, we discuss the situation where the dimension of $\mathcal{W}$ is finite. This result is employed in Section VII to provide the connection between the introduced Koopman learning problem and the EDMD method [34].

Theorem 4: Let the hypotheses of Theorem 3 hold, $w_1, \ldots, w_{n_{\mathrm{w}}} \in \mathcal{H}$ be linear independent vectors such that $\mathcal{W} = \operatorname{span}\{w_1, \ldots, w_{n_{\mathrm{w}}}\}$, matrix $\mathrm{P}_{\mathcal{W}}$ be defined as $\mathrm{P}_{\mathcal{W}} = [w_l(\mathrm{x}_{k-1})]_{k=1,l=1}^{n_{\mathrm{s}}, n_{\mathrm{w}}}$, and, W be the Gramian matrix of vectors $w_1, \ldots, w_{n_{\mathrm{w}}}$. Then, one can represent the unique solution of (25) as

$$\hat{\mathrm{K}}_{\mathcal{W}} = \sum_{j=1}^{n_{\mathrm{w}}} \sum_{l=1}^{n_{\mathrm{g}}} c_{jl} \, w_j \otimes g_l \tag{28}$$

where $\mathrm{C} = [c_{jl}]_{j=1,l=1}^{n_{\mathrm{w}}, n_{\mathrm{g}}}$ is the solution of the following optimization problem:

$$\min_{\mathrm{C} \in \mathbb{R}^{n_{\mathrm{w}} \times n_{\mathrm{g}}}} \ \|\mathrm{P}_{\mathcal{W}} \mathrm{C}\mathrm{G} - \mathrm{Y}\|_{\mathrm{F}}^2 + \lambda \|\mathrm{W}^{\frac{1}{2}} \mathrm{C}\mathrm{G}^{\frac{1}{2}}\|^2. \tag{29}$$

Proof: See Appendix B. ∎

Remark 4: From the proof of Theorem 4, one can see that $\mathrm{C} = \mathrm{W}^{-1} \mathrm{P}_{\mathcal{W}} \mathrm{A}$ and $\mathrm{W}_{\mathcal{V}} = \mathrm{P}_{\mathcal{W}} \mathrm{W}^{-1} \mathrm{P}_{\mathcal{W}}^{\mathsf{T}}$. These identities can be used as change-of-variable tricks for obtaining more tractable optimization problems.

Corollary 5: If $\mathcal{V} := \operatorname{span}\{v_1, \ldots, v_{n_{\mathrm{s}}}\} \subseteq \mathcal{W}$, then we have

$$\widetilde{\mathrm{K}}_{\mathcal{W}} = \hat{\mathrm{K}}_{\mathcal{W}} = \hat{\mathrm{K}}. \tag{30}$$

The approach discussed in Theorem 1 for learning the Koopman operator $\hat{\mathrm{K}}$ demands the knowledge of $v_1, \ldots, v_{n_\mathrm{s}}$. This issue can be addressed by Theorem 3 when $\Pi_{\mathcal{W}} v_1, \ldots, \Pi_{\mathcal{W}} v_{n_\mathrm{s}}$ are known. Note that this knowledge is not sufficient for the scheme proposed in Theorem 2 that approximates $\hat{\mathrm{K}}$ in $\mathcal{L}_{\mathcal{W}}$, where indeed, the knowledge of $v_1, \ldots, v_{n_\mathrm{s}}$ is again required to first solve problem (6), and then derive the approximation. On the other hand, in the case of finite dimensional space $\mathcal{W}$, when vectors $w_1, \ldots, w_{n_\mathrm{w}} \in \mathcal{H}$ are given such that $\mathcal{W} = \operatorname{span}\{w_1, \ldots, w_{n_\mathrm{w}}\}$, for solving the learning problem (25), it is enough to know $\{w_m(\mathrm{x}_{k-1})\}_{k=1,m=1}^{n_\mathrm{s},n_\mathrm{g}}$. In this situation, the knowledge of $v_1, \ldots, v_{n_\mathrm{s}}$ is not required. Also, from Corollary 5, we can see that if the set of vectors $w_1, \ldots, w_{n_\mathrm{w}}$ is rich enough in the sense that $\mathcal{V} \subseteq \mathcal{W} = \operatorname{span}\{w_1, \ldots, w_{n_\mathrm{w}}\}$, then the learning problems (6), (25), and (29) admit same solution. Hence, under the condition $\mathcal{V} \subseteq \mathcal{W} = \operatorname{span}\{w_1, \ldots, w_{n_\mathrm{w}}\}$, when $\{w_m(\mathrm{x}_{k-1})\}_{k=1,m=1}^{n_\mathrm{s},n_\mathrm{g}}$ is given, we can solve each of the above-mentioned learning problems without the knowledge of $v_1, \ldots, v_{n_\mathrm{s}}$.

VII. Connection to the EDMD

In this section, we consider the case where $\mathcal{G} := \operatorname{span}\{g_1, \ldots, g_{n_\mathrm{g}}\}$ is an invariant subspace of the learned Koopman operator, i.e., based on the notations introduced in Section IV, we need to have $\hat{\mathrm{K}} \in \mathcal{L}_{\mathcal{G}}$. Without loss of generality, we assume $g_1, \ldots, g_{n_\mathrm{g}}$ are linearly independent.

In the EDMD method, the Koopman operator is approximated by a finite-dimensional linear map $\mathrm{U} : \mathcal{G} \to \mathcal{G}$, and then, the observation data $\mathscr{D}$ is employed to estimate this map [34]. Since the dimension of $\mathcal{G}$ is finite, the map U admits a matrix representation in the basis $\{g_1, \ldots, g_{n_\mathrm{g}}\}$. More precisely, there exists matrix $\mathrm{M} \in \mathbb{R}^{n_\mathrm{g} \times n_\mathrm{g}}$ such that $\mathrm{U} g_l = \sum_{j=1}^{n_\mathrm{g}} [\mathrm{M}]_{(j,l)} g_j$, for $l = 1, \ldots, n_\mathrm{g}$. The matrix M is estimated by minimizing the empirical loss $\mathcal{E}_\mathrm{U}$ defined as

$$\mathcal{E}_\mathrm{U}(\mathrm{M}) := \|\mathrm{P}_{\mathcal{G}}\mathrm{M} - \mathrm{Y}\|_\mathrm{F}^2 = \sum_{k=1}^{n_\mathrm{s}} \sum_{l=1}^{n_\mathrm{g}} \left((\mathrm{U} g_l)(\mathrm{x}_{k-1}) - g_l(\mathrm{x}_k)\right)^2$$

where $\mathrm{P}_{\mathcal{G}} := [g_l(\mathrm{x}_{k-1})]_{k=1,l=1}^{n_\mathrm{s},\,n_\mathrm{g}}$, which is assumed to be full column rank, i.e., $\operatorname{rank}(\mathrm{P}_{\mathcal{G}}) = n_\mathrm{g}$. Accordingly, the empirical loss $\mathcal{E}_\mathrm{U}$ has a unique minimizer $\mathrm{M}^\star = \mathrm{P}_{\mathcal{G}}^\dagger \mathrm{Y}$, where $\mathrm{P}_{\mathcal{G}}^\dagger := (\mathrm{P}_{\mathcal{G}}^\mathsf{T}\mathrm{P}_{\mathcal{G}})^{-1}\mathrm{P}_{\mathcal{G}}^\mathsf{T}$ is the Moore–Penrose pseudoinverse of $\mathrm{P}_{\mathcal{G}}$.

Theorem 6: Define matrix $\mathrm{C}_{\mathrm{M}^\star}$ as $\mathrm{C}_{\mathrm{M}^\star} := \mathrm{M}^\star \mathrm{G}^{-1}$, and, let operator $\hat{\mathrm{K}}_\mathrm{U} \in \mathcal{L}(\mathcal{H})$ be defined as

$$\hat{\mathrm{K}}_\mathrm{U} = \sum_{j=1}^{n_\mathrm{g}} \sum_{l=1}^{n_\mathrm{g}} [\mathrm{C}_{\mathrm{M}^\star}]_{(j,l)}\, g_j \otimes g_l. \tag{31}$$

Then, $\hat{\mathrm{K}}_\mathrm{U}$ belongs to $\operatorname{argmin}_{\mathrm{K} \in \mathcal{L}_{\mathcal{G}}} \mathcal{E}(\mathrm{K})$, and, the EDMD map U coincides with the restriction of $\hat{\mathrm{K}}_\mathrm{U}$ to $\mathcal{G}$, i.e., $\mathrm{U} = \hat{\mathrm{K}}_\mathrm{U}\big|_{\mathcal{G}}$.

Proof: Similar to the proof of Theorem 3, one can show that $\mathcal{E}(\mathrm{K}) = \mathcal{E}(\Pi_{\mathcal{G}}\mathrm{K}\Pi_{\mathcal{G}})$, for each $\mathrm{K} \in \mathcal{L}_{\mathcal{G}}$. Accordingly, if $\min_{\mathrm{K} \in \mathcal{L}_{\mathcal{G}}} \mathcal{E}(\mathrm{K})$ admits a solution, it has also a solution in the set of operators $\mathcal{L}(\mathcal{G})$, which can be characterized as $\{\mathrm{K}_\mathrm{C} := \sum_{j=1}^{n_\mathrm{g}} \sum_{l=1}^{n_\mathrm{g}} c_{jl}\, g_j \otimes g_l \,|\, \mathrm{C} \in \mathbb{R}^{n_\mathrm{g} \times n_\mathrm{g}}\}$. Therefore, it is enough to consider the problem $\min_{\mathrm{K} \in \mathcal{L}(\mathcal{G})} \mathcal{E}(\mathrm{K})$. Given $\mathrm{C} \in \mathbb{R}^{n_\mathrm{g} \times n_\mathrm{g}}$, we define $\mathrm{M} = \mathrm{CG}$, where, based on same steps as in the proof of Theorem 4, we can show that $\mathcal{E}(\mathrm{K}_\mathrm{C}) = \mathcal{E}_\mathrm{U}(\mathrm{M})$. Hence, there is a bijection between the value and solutions of $\min_{\mathrm{K} \in \mathcal{L}(\mathcal{G})} \mathcal{E}(\mathrm{K})$ and $\min_{\mathrm{M} \in \mathbb{R}^{n_\mathrm{g} \times n_\mathrm{g}}} \mathcal{E}_\mathrm{U}(\mathrm{M})$. Since $\mathrm{M}^\star$ is a solution for the latter problem, then $\mathrm{K}_{\mathrm{C}_{\mathrm{M}^\star}}$, which coincides with $\hat{\mathrm{K}}_\mathrm{U}$, is a solution of $\min_{\mathrm{K} \in \mathcal{L}(\mathcal{G})} \mathcal{E}(\mathrm{K})$, which concludes the proof of first part. Due to (31), we have

$$\begin{aligned} \hat{\mathrm{K}}_\mathrm{U} g_i &= \textstyle\sum_{j=1}^{n_\mathrm{g}} \sum_{l=1}^{n_\mathrm{g}} [\mathrm{C}_{\mathrm{M}^\star}]_{(j,l)} g_j \langle g_l, g_i \rangle \\ &= \textstyle\sum_{j=1}^{n_\mathrm{g}} [\mathrm{C}_{\mathrm{M}^\star}\mathrm{G}]_{(j,l)} g_j = \sum_{j=1}^{n_\mathrm{g}} [\mathrm{M}^\star]_{(j,l)} g_j \end{aligned} \tag{32}$$

for each $i \in [n_\mathrm{g}]$, where the last equality is due to $\mathrm{M}^\star = \mathrm{C}_{\mathrm{M}^\star}\mathrm{G}$. ■

The following theorem demonstrates the connection between EDMD and the introduced learning problems.

Theorem 7: For $\lambda > 0$, let $\hat{\mathrm{K}}_{\mathcal{G},\lambda}$ be the unique solution of

$$\min_{\mathrm{K} \in \mathcal{L}_{\mathcal{G}}} \mathcal{J}_\lambda(\mathrm{K}) := \mathcal{E}(\mathrm{K}) + \lambda \|\mathrm{K}\|^2. \tag{33}$$

Then, $\lim_{\lambda \downarrow 0} \hat{\mathrm{K}}_{\mathcal{G},\lambda} = \hat{\mathrm{K}}_\mathrm{U}$ and $\lim_{\lambda \to \infty} \hat{\mathrm{K}}_{\mathcal{G},\lambda} = 0$, both in operator norm topology.

Proof: From Theorems 3 and 4, we know that, for each $\lambda > 0$, (33) admits a unique solution $\hat{\mathrm{K}}_{\mathcal{G},\lambda} = \sum_{j=1}^{n_\mathrm{g}} \sum_{l=1}^{n_\mathrm{g}} [\mathrm{C}_\lambda]_{(j,l)}\, g_j \otimes g_l$, where C_λ is defined as

$$\mathrm{C}_\lambda = \operatorname*{argmin}_{\mathrm{C} \in \mathbb{R}^{n_\mathrm{w} \times n_\mathrm{g}}} \|\mathrm{P}_{\mathcal{G}}\mathrm{CG} - \mathrm{Y}\|_\mathrm{F}^2 + \lambda \|\mathrm{G}^{\frac{1}{2}}\mathrm{CG}^{\frac{1}{2}}\|^2. \tag{34}$$

By the definition of C_λ, we have

$$\|\mathrm{P}_{\mathcal{G}}\mathrm{C}_\lambda\mathrm{G} - \mathrm{Y}\|_\mathrm{F}^2 \le \|\mathrm{P}_{\mathcal{G}}\mathrm{C}_\lambda\mathrm{G} - \mathrm{Y}\|_\mathrm{F}^2 + \lambda \|\mathrm{G}^{\frac{1}{2}}\mathrm{C}_\lambda\mathrm{G}^{\frac{1}{2}}\|^2 \le \|\mathrm{Y}\|_\mathrm{F}^2$$

which implies that $\|\mathrm{P}_{\mathcal{G}}\mathrm{C}_\lambda\mathrm{G}\|_\mathrm{F} \le 2\|\mathrm{Y}\|_\mathrm{F}$. Using Lemma 15 in Appendix H, we have

$$\begin{aligned} \|\mathrm{C}_\lambda\|_\mathrm{F} &= \|(\mathrm{P}_{\mathcal{G}}^\mathsf{T}\mathrm{P}_{\mathcal{G}})^{-1}\mathrm{P}_{\mathcal{G}}^\mathsf{T}(\mathrm{P}_{\mathcal{G}}\mathrm{C}_\lambda\mathrm{G})\mathrm{G}^{-1}\|_\mathrm{F} \\ &\le \|\mathrm{P}_{\mathcal{G}}^\dagger\| \|\mathrm{P}_{\mathcal{G}}\mathrm{C}_\lambda\mathrm{G}\|_\mathrm{F} \|\mathrm{G}^{-1}\| \le 2\|\mathrm{P}_{\mathcal{G}}^\dagger\| \|\mathrm{Y}\|_\mathrm{F} \|\mathrm{G}^{-1}\|. \end{aligned} \tag{35}$$

Accordingly, (34) in equivalent to the following program:

$$\operatorname*{argmin}_{\mathrm{C} \in \mathcal{M}} J_\lambda(\mathrm{C}) := \|\mathrm{P}_{\mathcal{G}}\mathrm{CG} - \mathrm{Y}\|_\mathrm{F}^2 + \lambda \|\mathrm{G}^{\frac{1}{2}}\mathrm{CG}^{\frac{1}{2}}\|^2 \tag{36}$$

where $\mathcal{M} := \{\mathrm{C} \in \mathbb{R}^{n_\mathrm{w} \times n_\mathrm{g}} \;\big|\; \|\mathrm{C}_\lambda\| \le 2\|\mathrm{P}_{\mathcal{G}}^\dagger\| \|\mathrm{Y}\|_\mathrm{F} \|\mathrm{G}^{-1}\|\}$, which is a convex and compact set. Let $\mathcal{C}_\lambda$ be the solution set of (36) for $\lambda \ge 0$, which is a singleton due to the strong convexity of the objective function. Moreover, we know that function J_λ is continuous with respect (C, λ). Hence, from Maximum Theorem [45], it follows that set-valued map $\lambda \mapsto \mathcal{C}_\lambda$ is upper hemicontinuous with nonempty and compact values, which implies that $\lim_{\lambda \downarrow 0} \mathcal{C}_\lambda = \mathcal{C}_0$, i.e., $\lim_{\lambda \downarrow 0} \mathrm{C}_\lambda = \mathrm{C}_{\mathrm{M}^\star}$. Accordingly, due to the structure of $\hat{\mathrm{K}}_\lambda$ and $\hat{\mathrm{K}}_\mathrm{U}$, we have $\lim_{\lambda \downarrow 0} \hat{\mathrm{K}}_{\mathcal{G},\lambda} = \hat{\mathrm{K}}_\mathrm{U}$ in operator norm topology. On the other hand, since $0 \in \mathcal{L}_{\mathcal{G}}$ is feasible for the problem and due to the definition of $\hat{\mathrm{K}}_{\mathcal{G},\lambda}$, we know that

$$\lambda \|\hat{\mathrm{K}}_{\mathcal{G},\lambda}\|^2 \le \mathcal{J}_\lambda(\hat{\mathrm{K}}_{\mathcal{G},\lambda}) \le \mathcal{J}_\lambda(0) = \|\mathrm{Y}\|_\mathrm{F}^2$$

which implies $\|\hat{\mathrm{K}}_{\mathcal{G},\lambda}\| \le \frac{1}{\sqrt{\lambda}} \|\mathrm{Y}\|_\mathrm{F}$. Therefore, we have $\lim_{\lambda \to \infty} \|\hat{\mathrm{K}}_{\mathcal{G},\lambda}\| = 0$, and subsequently, it follows that $\lim_{\lambda \to +\infty} \hat{\mathrm{K}}_{\mathcal{G},\lambda} = 0$, in operator norm topology. ■

Remark 5: For problem (6), one can see as $\lambda \to \infty$, the unique solution introduced in Theorem 1 converges to zero, in operator norm topology. The same claim holds for the unique solution of (25).

VIII. Generalized Representer Theorem and Cases of Interest

Considering the set of indices $\mathcal{I} \subseteq \{1,\dots,n_{\mathrm{s}}\} \times \{1,\dots,n_{\mathrm{g}}\}$ and function $\ell : \mathbb{R}^{|\mathcal{I}|} \times \mathbb{R}^{|\mathcal{I}|} \to \mathbb{R}_+$, the empirical loss can be generally defined as

$$\mathcal{E}_\ell(\mathrm{K}) = \ell\left([(\mathrm{K}g_l)(\mathrm{x}_{k-1})]_{(k,l)\in\mathcal{I}}, [y_{kl}]_{(k,l)\in\mathcal{I}}\right) \tag{37}$$

which is a convex function when $\ell(\cdot, \mathrm{Y}_\mathcal{I}) : \mathbb{R}^{|\mathcal{I}|} \to \mathbb{R}$ is convex for $\mathrm{Y}_\mathcal{I} := [y_{kl}]_{(k,l)\in\mathcal{I}} \in \mathbb{R}^{|\mathcal{I}|}$. Note that (37) generalizes various forms of empirical loss functions including the ones introduced in robust learning and statistics. Given a generic regularization function $\mathcal{R} : \mathcal{L}(\mathcal{H}) \to \mathbb{R}_+ \cup \{+\infty\}$ and $\lambda \in \mathbb{R}_+$, the learning problem for the Koopman operator can be defined in the most general form as

$$\begin{aligned} \min_{\mathrm{K}\in\mathcal{F}} \quad & \mathcal{J}_\ell(\mathrm{K}) := \mathcal{E}_\ell(\mathrm{K}) + \lambda\mathcal{R}(\mathrm{K}) \\ \text{s.t.} \quad & \mathrm{K} \in \mathcal{C} \end{aligned} \tag{38}$$

where $\mathcal{F}$ either denotes $\mathcal{L}(\mathcal{H})$, or $\mathcal{L}_\mathcal{W}$, for a closed subspace $\mathcal{W} \subseteq \mathcal{H}$, and $\mathcal{C}$ is a subset of $\mathcal{H}$. In the following, we determine when learning problem (38) is tractable, i.e., we provide suitable conditions under which a representer theorem holds for (38).

Notational conventions: Define $\overline{\mathcal{R}} : \mathcal{L}(\mathcal{H}) \to \mathbb{R} \cup \{+\infty\}$ as $\overline{\mathcal{R}} := \lambda\mathcal{R} + \delta_\mathcal{C}$, and let $\Pi_\mathcal{G}$ be the projection operator on $\mathcal{G} = \mathrm{span}\{g_1,\dots,g_{n_{\mathrm{g}}}\}$. For brevity, we unify the notations for both cases of $\mathcal{F}$. Indeed, when $\mathcal{F}$ is $\mathcal{L}_\mathcal{W}$, we set n_{z} as n_{s}, z_k as $z_k = \Pi_\mathcal{W} v_k$, for $k = 1,\dots,n_{\mathrm{z}}$, and $\mathcal{Z}$ as the subspace $\mathcal{Z} = \mathrm{span}\{z_1,\dots,z_{n_{\mathrm{z}}}\}$. Moreover, we denote by $\Pi_\mathcal{Z}$ and Z, respectively, as the projection operator on $\mathcal{Z}$ and the Gramian matrix of vectors $z_1,\dots,z_{n_{\mathrm{z}}}$. By abuse of notation, we use the same letters for n_{s}, $v_1,\dots,v_{n_{\mathrm{s}}}$, $\mathcal{V}$, $\Pi_\mathcal{V}$, and, V, for the case $\mathcal{F} = \mathcal{L}(\mathcal{H})$. Also, in order to provide arguments analogous to the case of finite dimensional subspace $\mathcal{W} = \mathrm{span}\{w_1,\dots,w_{n_{\mathrm{w}}}\}$ as in Theorem 4, we adopt the same notational convention for n_{w}, $w_1,\dots,w_{n_{\mathrm{w}}}$, $\mathcal{W}$, $\Pi_\mathcal{W}$, and, W. To provide a generalized representer theorem for (38), we need the following assumption.

Assumption 2: For any $\mathrm{S} \in \mathcal{L}(\mathcal{H})$, we have

$$\overline{\mathcal{R}}(\Pi_\mathcal{Z}\mathrm{S}\Pi_\mathcal{G}) \le \overline{\mathcal{R}}(\mathrm{S}). \tag{39}$$

Based on the discussions in Section IV for learning the Koopman operator with Tikhonov regularization, one can see that the property (39) holds for (6) and plays a crucial role in deriving the presented results. The following theorem provides analogous result for (38). More precisely, we show that the learning problem formulated as the infinite-dimensional program (38) admits an exact finite-dimensional convex reformulation.

Theorem 8 (Generalized Representer Theorem): Let Assumptions 1 and 2 hold, and $v_1,\dots,v_{n_{\mathrm{s}}} \in \mathcal{H}$ be the vectors defined in Theorem 1.

i) Suppose that the optimization problem (38) admits a solution. Then, (38) has a solution in the following form

$$\hat{\mathrm{K}} = \sum_{k=1}^{n_{\mathrm{z}}} \sum_{l=1}^{n_{\mathrm{g}}} a_{kl}\, z_k \otimes g_l \tag{40}$$

where $a_{kl} \in \mathbb{R}$, for $k \in [n_{\mathrm{z}}]$ and $l \in [n_{\mathrm{g}}]$.

ii) When $\mathcal{D} := \mathcal{F} \cap \mathrm{dom}(\mathcal{R}) \cap \mathcal{C}$ is a nonempty, closed, and convex set, $\ell(\cdot, \mathrm{Y}_\mathcal{I}) : \mathbb{R}^{|\mathcal{I}|} \to \mathbb{R}$ is a convex function for each $\mathrm{Y}_\mathcal{I} := [y_{kl}]_{(k,l)\in\mathcal{I}} \in \mathbb{R}^{|\mathcal{I}|}$, $\mathcal{R}$ is convex and lower semicontinuous, and $\overline{\mathcal{R}}$ is coercive, then, (38) admits at least one solution with parametric representation (40). Additionally, if $\mathcal{R}$ is strictly convex on $\mathcal{D}$, then the solution of (38) is unique and admits parametric form in (40).

Proof: Define function $\overline{\mathcal{J}}_\ell : \mathcal{L}(\mathcal{H}) \to \mathbb{R} \cup \{+\infty\}$ as

$$\overline{\mathcal{J}}_\ell(\mathrm{K}) := \mathcal{E}_\ell(\mathrm{K}) + \lambda\mathcal{R}(\mathrm{K}) + \delta_\mathcal{C}(\mathrm{K}) + \delta_\mathcal{F}(\mathrm{K}) \tag{41}$$

for each $\mathrm{K} \in \mathcal{L}(\mathcal{H})$. One can see that the learning problem (38) is equivalent to $\min_{\mathrm{K}\in\mathcal{L}(\mathcal{H})} \overline{\mathcal{J}}_\ell(\mathrm{K})$. Let S be a solution of (38), which is clearly a solution for this optimization problem as well. Consider the case $\mathcal{F}$ is $\mathcal{L}(\mathcal{H})$ and let operator $\hat{\mathrm{K}}$ be defined as $\hat{\mathrm{K}} = \Pi_\mathcal{V}\mathrm{S}\Pi_\mathcal{G}$. In this case, the last term in (41) is zero for S and $\hat{\mathrm{K}}$. Moreover, similar to the proof of Theorem 1, one can show $(\hat{\mathrm{K}}g_l)(\mathrm{x}_{k-1}) = (\mathrm{S}g_l)(\mathrm{x}_{k-1})$, for each $k \in [n_{\mathrm{s}}]$ and $l \in [n_{\mathrm{g}}]$. Hence, by the definition of $\mathcal{E}_\ell$ in (37), we see that $\mathcal{E}_\ell(\hat{\mathrm{K}}) = \mathcal{E}_\ell(\mathrm{S})$. Due to Assumption 2, we have

$$\overline{\mathcal{J}}_\ell(\hat{\mathrm{K}}) = \mathcal{E}_\ell(\hat{\mathrm{K}}) + \overline{\mathcal{R}}(\hat{\mathrm{K}}) \le \mathcal{E}_\ell(\mathrm{S}) + \overline{\mathcal{R}}(\mathrm{S}) = \overline{\mathcal{J}}_\ell(\mathrm{S}). \tag{42}$$

Thus, $\hat{\mathrm{K}}$ is a solution of (38) as well, and, we have $\overline{\mathcal{J}}_\ell(\hat{\mathrm{K}}) = \overline{\mathcal{J}}_\ell(\mathrm{S})$. From linearity of $\hat{\mathrm{K}}$, it follows that $\hat{\mathrm{K}}$ admits the parametric form in (40). For the case $\mathcal{F} = \mathcal{L}_\mathcal{W}$, define operator $\hat{\mathrm{K}}_\mathcal{W}$ as $\hat{\mathrm{K}}_\mathcal{W} = \Pi_\mathcal{W}\mathrm{S}\Pi_\mathcal{G}$. By similar arguments, we can show that $\hat{\mathrm{K}}_\mathcal{W}$ is a solution of (38) for which we have the given parametric representation. This concludes the proof for part i).

Due to the convexity property of ℓ, we know that $\ell(\cdot, \mathrm{Y}_\mathcal{I}) : \mathbb{R}^{|\mathcal{I}|} \to \mathbb{R}$ is a continuous function, for $\mathrm{Y}_\mathcal{I} = [y_{kl}]_{(k,l)\in\mathcal{I}} \in \mathbb{R}^{|\mathcal{I}|}$. Furthermore, we know that

$$\mathcal{E}_\ell(\mathrm{K}) = \ell\left([\langle \mathrm{K}g_l, \mathrm{v}_k\rangle]_{(k,l)\in\mathcal{I}}, [y_{kl}]_{(k,l)\in\mathcal{I}}\right) \tag{43}$$

which implies that $\mathcal{E}_\ell : \mathcal{H} \to \mathbb{R}_+$ is a proper, convex and continuous function. Moreover, $\mathcal{R}$ is convex and lower semicontinuous. Note that we have $\mathrm{dom}(\overline{\mathcal{J}}_\ell) = \mathcal{F} \cap \mathrm{dom}(\mathcal{R}) \cap \mathcal{C} = \mathcal{D}$, which is a nonempty, closed and convex set. Thus, $\overline{\mathcal{J}}_\ell$ is a proper function, and also $\delta_\mathcal{D}$ is proper, convex, and lower semicontinuous. One can see that $\overline{\mathcal{J}}_\ell(\mathrm{K}) = \mathcal{E}_\ell(\mathrm{K}) + \mathcal{R}(\mathrm{K}) + \delta_\mathcal{D}(\mathrm{K})$, for each $\mathrm{K} \in \mathcal{H}$. Therefore, $\overline{\mathcal{J}}_\ell$ is proper, convex, and lower semicontinuous. Moreover, it is strictly convex, when $\mathcal{R}$ is a strictly convex function. From nonnegativity of $\mathcal{E}_\ell$ and $\delta_\mathcal{D}$, and, due to $\overline{\mathcal{J}}_\ell(\mathrm{K}) = \mathcal{E}_\ell(\mathrm{K}) + \overline{\mathcal{R}}(\mathrm{K}) + \delta_\mathcal{D}(\mathrm{K})$, for each $\mathrm{K} \in \mathcal{H}$, it follows that $\overline{\mathcal{J}}_\ell$ is coercive. Therefore, (38) admits at least one solution, which is unique when $\mathcal{R}$ is strictly convex [43]. This concludes the proof for part ii). ■

Remark 6: Let $\mathcal{D} := \mathcal{F} \cap \mathrm{dom}(\mathcal{R}) \cap \mathcal{C}$, and, define $\mathcal{J}_\ell : \mathcal{D} \to \mathbb{R}$ as the restriction of function $\overline{\mathcal{J}}_\ell$ introduced (41) to $\mathcal{D}$, i.e., $\mathcal{J}_\ell = \overline{\mathcal{J}}_\ell|_\mathcal{D}$. Due to *variational principle* [42], if $\mathcal{D}$ is a nonempty weakly sequentially closed set, and $\mathcal{J}_\ell : \mathcal{D} \to \mathbb{R}$ is weakly sequentially lower semicontinuous and coercive, then there exists $\mathrm{K} \in \mathcal{D}$ such that $\mathcal{J}_\ell(\mathrm{K}) = \inf_{\mathrm{S}\in\mathcal{D}} \mathcal{J}_\ell(\mathrm{S})$, i.e., (38) has a solution. Note that here no convexity assumption is required, and hence, these conditions are more general.

Theorem 9:

i) Let $\lambda \in \mathbb{R}_+$, $\mathcal{R} : \mathcal{L}(\mathcal{H}) \to \mathbb{R}_+ \cup \{+\infty\}$ be a regularization function, and, $\mathcal{C}_\alpha$ be a given subset of $\mathcal{H}$, for each $\alpha \in \mathcal{A}$, where $\mathcal{A}$ is the corresponding set of indices. Define $\mathcal{C} := \cap_{\alpha\in\mathcal{A}} \mathcal{C}_\alpha$ and $\overline{\mathcal{R}}_\alpha := \lambda\mathcal{R} + \delta_{\mathcal{C}_\alpha}$, for $\alpha \in \mathcal{A}$. If Assumption 2 is satisfied by $\overline{\mathcal{R}}_\alpha$, for each $\alpha \in \mathcal{A}$, then $\overline{\mathcal{R}} := \lambda\mathcal{R} + \delta_\mathcal{C}$ also satisfies Assumption 2.

ii) Let $\mathcal{R}_1, \ldots, \mathcal{R}_m : \mathcal{L}(\mathcal{H}) \to \mathbb{R}_+ \cup \{+\infty\}$ be given regularization functions, $\lambda_1, \ldots, \lambda_m \in \mathbb{R}_+$ and $\mathcal{C}$ be a subset of $\mathcal{H}$. If $\overline{\mathcal{R}}_i = \lambda_i \mathcal{R}_i + \delta_{\mathcal{C}}$ satisfies Assumption 2, for $i \in [m]$, then $\overline{\mathcal{R}} := \sum_{i=1}^m \lambda_i \mathcal{R}_i + \delta_{\mathcal{C}}$ also satisfies Assumption 2.

Proof:

i) Let S be an arbitrary element in $\mathcal{L}(\mathcal{H})$. If $\mathrm{S} \notin \mathcal{C}$ or $\lambda\mathcal{R}(\mathrm{S}) = +\infty$, then (39) holds trivially. Hence, we assume that $\lambda\mathcal{R}(\mathrm{S})$ is finite and $\mathrm{S} \in \mathcal{C}$. Therefore, for each α, we know that $\mathrm{S} \in \mathcal{C}_\alpha$, and subsequently, $\delta_{\mathcal{C}_\alpha}(\mathrm{S}) = 0$. Since Assumption 2 holds for $\overline{\mathcal{R}}_\alpha$, we have

$$\lambda\mathcal{R}(\Pi_{\mathcal{Z}}\mathrm{S}\Pi_{\mathcal{G}}) + \delta_{\mathcal{C}_\alpha}(\Pi_{\mathcal{Z}}\mathrm{S}\Pi_{\mathcal{G}}) \leq \lambda\mathcal{R}(\mathrm{S}) + \delta_{\mathcal{C}_\alpha}(\mathrm{S}) = \lambda\mathcal{R}(\mathrm{S})$$

which implies that $\Pi_{\mathcal{Z}}\mathrm{S}\Pi_{\mathcal{G}} \in \mathcal{C}_\alpha$ and $\lambda\mathcal{R}(\Pi_{\mathcal{Z}}\mathrm{S}\Pi_{\mathcal{G}}) \leq \lambda\mathcal{R}(\Pi_{\mathcal{Z}}\mathrm{S}\Pi_{\mathcal{G}})$. Therefore, we have $\Pi_{\mathcal{Z}}\mathrm{S}\Pi_{\mathcal{G}} \in \mathcal{C}$, and $\delta_{\mathcal{C}}(\Pi_{\mathcal{Z}}\mathrm{S}\Pi_{\mathcal{G}}) = 0$. Subsequently, it follows that $\overline{\mathcal{R}}(\Pi_{\mathcal{Z}}\mathrm{S}\Pi_{\mathcal{G}}) \leq \overline{\mathcal{R}}(\mathrm{S})$. Hence, Assumption 2 is satisfied by $\overline{\mathcal{R}}$.

ii) It is easy to check that

$$\overline{\mathcal{R}}(\mathrm{S}) = \sum_{i=1}^m \lambda_i \mathcal{R}_i(\mathrm{S}) + \delta_{\mathcal{C}}(\mathrm{S}) = \sum_{i=1}^m (\lambda_i \mathcal{R}_i(\mathrm{S}) + \delta_{\mathcal{C}}(\mathrm{S}))$$

for any $\mathrm{S} \in \mathcal{L}(\mathcal{H})$. Thus, the proof is implied directly. ■

Remark 7: Theorem 9 allows utilizing the result of Theorem 8 for a wide range of interest case, where different combinations of constraints and regularization terms are considered together in the Koopman learning problem (38).

In the following, we provide applications of the theorems above.

A. Learning Koopman Operator With Frobenius Norm Regularization

In Section IV, the introduced regularization function is based on the operator norm. On the other hand, one may propose employing Hilbert–Schmidt or Frobenius norm of the Koopman operator to define the regularization term, i.e., $\mathcal{R} : \mathcal{L}(\mathcal{H}) \to \mathbb{R}_+ \cup \{+\infty\}$ is defined as $\mathcal{R}(\mathrm{K}) := \|\mathrm{K}\|_{\mathrm{F}}^2$, for $\mathrm{K} \in \mathcal{L}(\mathcal{H})$. Thus, the learning problem is formulated as

$$\min_{\mathrm{K} \in \mathcal{F}} \ \mathcal{E}(\mathrm{K}) + \lambda \|\mathrm{K}\|_{\mathrm{F}}^2 \tag{44}$$

where $\mathcal{E}(\mathrm{K})$ is the empirical loss defined in (4). Before proceeding further, we recall that $\|\mathrm{S}\|_{\mathrm{F}}^2 = \mathrm{tr}(\mathrm{S}^*\mathrm{S})$, for each $\mathrm{S} \in \mathcal{L}(\mathcal{H})$. Moreover, given an orthonormal basis $\{b_k\}_{k=1}^\infty$ for $\mathcal{H}$, the trace of operator $\mathrm{S} \in \mathcal{L}_{\mathcal{W}}$ is defined as

$$\mathrm{tr}(\mathrm{S}) := \sum_{k=1}^\infty \langle b_k, \mathrm{S} b_k \rangle \tag{45}$$

when the summation converges [42]. Based on (45), we have

$$\mathcal{R}(\mathrm{K}) = \|\mathrm{K}\|_{\mathrm{F}}^2 = \sum_{k=1}^\infty \langle b_k, \mathrm{K}^*\mathrm{K} b_k \rangle = \sum_{k=1}^\infty \|\mathrm{K} b_k\|^2. \tag{46}$$

Note that the left-hand sides of (45) and (46) are independent of the choice of orthonormal basis $\{b_k\}_{k=1}^\infty$.

The following theorem characterizes the solution of the infinite-dimensional learning problem (44) through obtaining an exact finite-dimensional quadratic program reformulation.

Theorem 10: Let Assumption 1 hold and $\lambda > 0$. Then, the optimization problem (44) has a *unique* solution with parametric representation as in (40), where $\mathrm{A} = [a_{kl}]_{k=1,l=1}^{n_{\mathrm{z}},\, n_{\mathrm{g}}} \in \mathbb{R}^{n_{\mathrm{z}} \times n_{\mathrm{g}}}$ is the solution of the following quadratic program

$$\min_{\mathrm{A} \in \mathbb{R}^{n_{\mathrm{z}} \times n_{\mathrm{g}}}} \ \|\mathrm{ZAG} - \mathrm{Y}\|_{\mathrm{F}}^2 \ + \ \lambda \|\mathrm{Z}^{\frac{1}{2}} \mathrm{A} \mathrm{G}^{\frac{1}{2}}\|_{\mathrm{F}}^2. \tag{47}$$

Proof: See Appendix C. ■

Define $J_{\mathrm{F}} : \mathbb{R}^{n_{\mathrm{z}} \times n_{\mathrm{g}}} \to \mathbb{R}_+$ as the objective function in (47), i.e., for any $\mathrm{A} \in \mathbb{R}^{n_{\mathrm{z}} \times n_{\mathrm{g}}}$, we have

$$J_{\mathrm{F}}(\mathrm{A}) = \|\mathrm{ZAG} - \mathrm{Y}\|_{\mathrm{F}}^2 \ + \ \lambda \|\mathrm{Z}^{\frac{1}{2}} \mathrm{A} \mathrm{G}^{\frac{1}{2}}\|_{\mathrm{F}}^2. \tag{48}$$

For the first derivative of J_{F}, one can easily see that

$$\frac{1}{2} \nabla_{\mathrm{A}} J_{\mathrm{F}}(\mathrm{A}) = \mathrm{Z}^2 \mathrm{A} \mathrm{G}^2 + \lambda \mathrm{ZAG} - \mathrm{ZYG}. \tag{49}$$

To solve (47), we can use the first order necessary condition $\nabla_{\mathrm{A}} J_{\mathrm{F}}(\mathrm{A}) = 0$, which is a linear system of equation with respect to A. Indeed, $\nabla_{\mathrm{A}} J_{\mathrm{F}}(\mathrm{A}) = 0$ is a *generalized Sylvester equation*, which can be solved efficiently. More precisely, $\nabla_{\mathrm{A}} J_{\mathrm{F}}(\mathrm{A}) = 0$ has a closed form solution in terms of λ, matrix Y, and the SVD of matrices G and Z.

Remark 8: According to the discussion above, one can see that employing Frobenius norm for the regularization leads to less computationally demanding problem compared to the case of regularization with operator norm. Hence, one may prefer learning Koopman operator based on (44) rather than (6).

Remark 9: For $\lambda > 0$ and $\mathcal{F} = \mathcal{L}_{\mathcal{G}}$, let $\hat{\mathrm{K}}_{\mathcal{G},\lambda}$ denote the unique solution of learning problem (44). Similar to Theorem 7, one can show that $\lim_{\lambda \downarrow 0} \hat{\mathrm{K}}_{\mathcal{G},\lambda} = \hat{\mathrm{K}}_{\mathrm{U}}$ and $\lim_{\lambda \to \infty} \hat{\mathrm{K}}_{\mathcal{G},\lambda} = 0$, in the Frobenius norm and the operator norm topologies.

B. Learning Koopman Operator With Rank Constraint

Learning a low-rank operator can be of interest when a reduced version of the Koopman operator is desired, e.g., for the model reduction of the system [35], [46]. Thus, one may introduce a rank constraint in the learning problem as

$$\mathcal{C} := \left\{ \mathrm{S} \in \mathcal{L}(\mathcal{H}) \,\middle|\, \mathrm{rank}(\mathrm{S}) := \dim(\mathrm{S}(\mathcal{H})) \leq r \right\} \tag{50}$$

where $r \in \mathbb{Z}_+$ is a given bound on the rank of Koopman operator. Accordingly, the resulting learning problem is

$$\begin{array}{ll} \min\limits_{\mathrm{K} \in \mathcal{F}} & \mathcal{E}(\mathrm{K}) \\ \text{s.t.} & \mathrm{K} \in \mathcal{C} \end{array} \tag{51}$$

where $\mathcal{E}(\mathrm{K})$ is the empirical loss defined in (4). The following theorem says that for the infinite-dimensional program (51), there exists an exact equivalent finite-dimensional convex reformulation that can be solved efficiently to obtain the solution of (51).

Theorem 11: Under Assumption 1, if the optimization problem (51) admits a solution, it has a solution with parametric form given in (40), where $\mathrm{A} = [a_{kl}]_{k=1,l=1}^{n_{\mathrm{z}},\, n_{\mathrm{g}}} \in \mathbb{R}^{n_{\mathrm{z}} \times n_{\mathrm{g}}}$ is the solution of the following problem:

$$\begin{array}{ll} \min\limits_{\mathrm{A} \in \mathbb{R}^{n_{\mathrm{z}} \times n_{\mathrm{g}}}} & \|\mathrm{ZAG} - \mathrm{Y}\|_{\mathrm{F}}^2 \\ \text{s.t.} & \mathrm{rank}(\mathrm{ZAG}) \leq r. \end{array} \tag{52}$$

Proof: See Appendix D. ∎

Remark 10: The conditions in Theorem 8 or Remark 6 provided for guaranteeing the existence or uniqueness of the solution are not satisfied for the learning problem (51). More precisely, function $\overline{\mathcal{R}}$ is not coercive here.

Using change-of-variable $\mathrm{B} = \mathrm{ZAG}$, the optimization problem (52) can be modified to

$$\begin{aligned} \min_{\mathrm{B}\in\mathbb{R}^{n_\mathrm{z}\times n_\mathrm{g}}} \quad & \|\mathrm{B}-\mathrm{Y}\|_\mathrm{F}^2 \\ \text{s.t.} \quad & \mathrm{rank}(\mathrm{B}) \le \tilde{r} \end{aligned} \tag{53}$$

where $\tilde{r}$ is defined as $\tilde{r} = \min\{r, \mathrm{rank}(\mathrm{G}), \mathrm{rank}(\mathrm{Z})\}$. Following this, one can use Eckart–Young–Mirsky theorem [47] to solve (53). More precisely, let $\mathrm{Y} = \underline{\mathrm{U}}\Sigma\overline{\mathrm{U}}^\mathsf{T}$ be the SVD of matrix Y, $\Sigma_1 = \mathrm{diag}(\sigma_1, \ldots, \sigma_{\tilde{r}})$ be the diagonal matrix containing first $\tilde{r}$ largest singular values of Y, and, $\underline{\mathrm{U}}_1$ and $\overline{\mathrm{U}}_1$ be, respectively, the matrices containing first $\tilde{r}$ columns of $\underline{\mathrm{U}}$ and $\overline{\mathrm{U}}$. Then, $\mathrm{B} = \underline{\mathrm{U}}\Sigma_1\overline{\mathrm{U}}^\mathsf{T}$ solves (53), and subsequently, we can obtain A by solving $\mathrm{ZAG} = \underline{\mathrm{U}}\Sigma_1\overline{\mathrm{U}}^\mathsf{T}$ for A.

C. Learning Koopman Operator With Nuclear Norm Regularization

In various learning problems, like collaborative filtering, nuclear norm regularization is employed to penalize the complexity of the model [48]. Indeed, the nuclear norm is interpreted as a convex relaxation of rank [49]. Similar to the matrices, the nuclear norm of operator $\mathrm{S} \in \mathcal{L}(\mathcal{H})$ is defined as $\|\mathrm{S}\|_* = \mathrm{tr}(|\mathrm{S}|)$, where $|\mathrm{S}|$ denotes the square root of S, i.e., $|\mathrm{S}|$ is a nonnegative operator such that $|\mathrm{S}|^2 = \mathrm{S}^*\mathrm{S}$ [42]. Thus, due to (45), given an orthonormal basis $\{b_k\}_{k=1}^\infty$, we have

$$\|\mathrm{S}\|_* := \sum_{k=1}^{\infty} \langle b_k, |\mathrm{S}| b_k\rangle \tag{54}$$

when the summation converges [42]. Also, it is known that

$$\|\mathrm{K}\|_* = \sup\left\{|\mathrm{tr}(\mathrm{CK})| \;\middle|\; \mathrm{C}\in\mathcal{K}(\mathcal{H}), \|\mathrm{C}\|\le 1\right\} \tag{55}$$

where $\mathcal{K}(\mathcal{H})$ denotes the set of compact operators on $\mathcal{H}$ [42]. Considering nuclear norm of the Koopman operator as the regularization term, we have the following learning problem:

$$\min_{\mathrm{K}\in\mathcal{F}} \; \mathcal{E}(\mathrm{K}) + \lambda\|\mathrm{K}\|_* \tag{56}$$

where $\lambda > 0$ and $\mathcal{E}(\mathrm{K})$ is the empirical loss defined in (4). It can be shown that the infinite-dimensional learning problem (56) can be reformulated precisely to a finite-dimensional convex optimization.

Theorem 12: Under Assumption 1, the optimization problem (56) admits a solution $\hat{\mathrm{K}}$ with parametric form (40), where $\mathrm{A} = [a_{kl}]_{k=1,l=1}^{n_\mathrm{z},\,n_\mathrm{g}}$ is the solution of following convex program:

$$\min_{\mathrm{A}\in\mathbb{R}^{n_\mathrm{z}\times n_\mathrm{g}}} \; \|\mathrm{ZAG}-\mathrm{Y}\|_\mathrm{F}^2 + \lambda\|\mathrm{Z}^{\frac{1}{2}}\mathrm{AG}^{\frac{1}{2}}\|_* \,. \tag{57}$$

Proof: See Appendix E. ∎

Remark 11: The regularization function $\mathcal{R}(\cdot) := \|\cdot\|_*$, employed in (56), is not strictly convex. Therefore, Theorem 8 cannot guarantee the uniqueness of the solution. However, if we only consider the minimum norm solution of (56), then it is unique and attains the parametric form (40).

Remark 12: Let $\hat{\mathrm{K}}_{\mathcal{G},\lambda}$ be the unique minimum norm solution of the learning problem (56) with $\lambda > 0$ and $\mathcal{F} = \mathcal{L}_\mathcal{G}$. Similar to Theorem 7, we can show that $\lim_{\lambda\downarrow 0}\hat{\mathrm{K}}_{\mathcal{G},\lambda} = \hat{\mathrm{K}}_\mathrm{U}$, and $\lim_{\lambda\to\infty}\hat{\mathrm{K}}_{\mathcal{G},\lambda} = 0$, in nuclear norm topology, which implies convergence in Frobenius norm and operator norm as well.

D. Learning Stable Koopman Operator

Let $\mathrm{x}_\mathrm{eq} = [x_{\mathrm{eq},1}, \ldots, x_{\mathrm{eq},n}]^\mathsf{T} \in \mathcal{X}$ be an equilibrium point for dynamics (1). In this section, we assume the Hilbert space of observables is an RKHS $\mathcal{H}_\Bbbk$ endowed with kernel $\Bbbk : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ for which we have

$$\Bbbk(\mathrm{x}, \mathrm{x}_\mathrm{eq}) = \Bbbk(\mathrm{x}_\mathrm{eq}, \mathrm{x}) = 0, \qquad \forall \mathrm{x}\in\mathcal{X}. \tag{58}$$

Indeed, given kernel $\Bbbh : \mathcal{X}\times\mathcal{X}\to\mathbb{R}$, we can define $\Bbbk$ as

$$\Bbbk(\mathrm{x},\mathrm{y}) = \Bbbh(\mathrm{x},\mathrm{y}) - \Bbbh(\mathrm{x},\mathrm{x}_\mathrm{eq}) - \Bbbh(\mathrm{x}_\mathrm{eq},\mathrm{y}) + \Bbbh(\mathrm{x}_\mathrm{eq},\mathrm{x}_\mathrm{eq})$$

for each $\mathrm{x},\mathrm{y}\in\mathcal{X}$. Then, one can see that $\Bbbk$ is a positive definite kernel satisfying (58). Note that when $\Bbbk$ has this property, for each observable $g\in\mathcal{H}_\Bbbk$, we have

$$g(\mathrm{x}_\mathrm{eq}) = \langle \Bbbk(\mathrm{x}_\mathrm{eq}, \cdot), g\rangle = 0. \tag{59}$$

Assumption 3: There exist $h_1, \ldots, h_{n_\mathrm{x}} \in \mathcal{H}_\Bbbk$ and positive scalars $L_1, \ldots, L_{n_\mathrm{x}}, \alpha_1, \ldots, \alpha_{n_\mathrm{x}}$ such that, for each $\mathrm{x} = [x_1, \ldots, x_{n_\mathrm{x}}]^\mathsf{T} \in \mathcal{X}$, we have

$$|h_j(\mathrm{x})| \ge L_j|x_j - x_{\mathrm{eq},j}|^{\alpha_j}, \qquad \forall j\in\{1,\ldots,n_\mathrm{x}\}. \tag{60}$$

For example, if quadratic function $h(\mathrm{x}) = \|\mathrm{x}-\mathrm{x}_\mathrm{eq}\|^2$ belongs to $\mathcal{H}_\Bbbk$, then Assumption 3 is satisfied. More precisely, (60) holds for $h_j = h$, $L_j = 1$, and $\alpha_j = 2$, for $j\in[n_\mathrm{x}]$.

Theorem 13: Let Assumption 3 hold for $\mathcal{H}_\Bbbk$, $\sup_{\mathrm{x}\in\mathcal{X}}\Bbbk(\mathrm{x},\mathrm{x}) < \infty$, and there exist $\varepsilon > 0$ such that $\|\mathrm{K}\| \le 1-\varepsilon$. Then, x_eq is a globally stable equilibrium point.

Proof: See Appendix F. ∎

Motivated by Theorem 13, we can include in the learning problem the side-information on the stability of equilibrium point (38) as

$$\begin{aligned} \min_{\mathrm{K}\in\mathcal{F}} \quad & \mathcal{E}_\ell(\mathrm{K}) + \lambda\mathcal{R}(\mathrm{K}) \\ \text{s.t.} \quad & \mathrm{K}\in\mathcal{C} \\ & \|\mathrm{K}\| \le 1-\varepsilon \end{aligned} \tag{61}$$

with $\varepsilon > 0$. Similar to before, we have the following theorem.

Theorem 14: Under the hypotheses of Theorem 8, the existence, the uniqueness, and the parametric representation (40) are implied for the solution of learning problem (61).

Proof: See Appendix G. ∎

IX. Numerical Experiments and Examples

In this section, we provide numerical examples elaborating the presented results. Throughout this section, the Hilbert space of observables is specified as RKHS $\mathcal{H}_\Bbbk$ with kernel $\Bbbk : \mathcal{X}\times\mathcal{X}\to\mathbb{R}$, and observables $\{g_l\}_{l\in[n_\mathrm{g}]}$ are defined according to Remark 2, i.e., $g_l(\cdot) := \Bbbk(\mathrm{p}_l, \cdot)$, for $l\in[n_\mathrm{g}]$, where $\mathcal{P} := \{\mathrm{p}_l\}_{l=1}^{n_\mathrm{g}}$ is a finite set of points in $\mathcal{X}$. Furthermore, for tuning hyperparameters

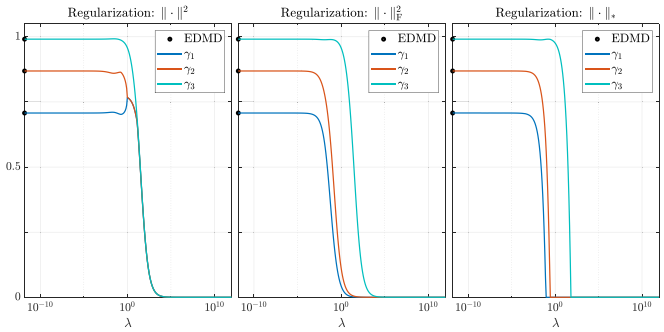

Fig. 2. Magnitude of eigenvalues derived from the proposed scheme converge to the ones obtained from EDMD as the weight of regularization, λ, goes to 0. We can see that convergence has different behavior depending on the choice of regularization.

such as the regularization weight, we employ a cross-validation scheme implemented using a Bayesian optimization heuristic.

Example 1: The connection between EDMD and learning problems (6), (44), and, (56) is discussed respectively in Theorem 7, Remark 9, and Remark 12. To illustrates this feature, we consider the following stable nonlinear dynamics

$$\begin{aligned} x_{1,k+1} &= \mu_1 x_{1,k} \\ x_{2,k+1} &= \mu_2 x_{2,k} + \left(\mu_1^2 - \mu_2\right) x_{1,k}^2 \end{aligned} \tag{62}$$

where $\mu_1 = 0.95$ and $\mu_2 = 0.75$ [30].

Consider Gaussian kernel $\mathbb{k}$ defined as

$$\mathbb{k}(\mathrm{x},\mathrm{y}) = \mathrm{e}^{-\frac{1}{2\ell_{\mathbb{k}}^2}\|\mathrm{x}-\mathrm{y}\|^2} \quad \forall, \mathrm{x}, \mathrm{y} \in \mathbb{R}^2 \tag{63}$$

where $\ell_{\mathbb{k}} = 1$. The system is initialized at $\mathrm{x}_0 = [1,0]^\mathsf{T}$, and then generated a trajectory of length $n_{\mathrm{s}} = 60$. Let g_1, g_2, and g_3 be three observables defined, respectively, as section of kernel $\mathbb{k}$ at $\mathrm{p}_1 = [1,0]^\mathsf{T}$, $\mathrm{p}_2 = [1,1]^\mathsf{T}$, and $\mathrm{p}_3 = [0,1]^\mathsf{T}$. Suppose that we have the values of these observables along the trajectory of the system. Given this data, we can apply the EDMD method, and also, the scheme introduced in Theorem 4, for different values of λ and, $w_l := g_l$, for $l = 1, 2, 3$. Each of these methods provides an estimation of the eigenvalues of the Koopman operator. From Theorem 7, we expect that as the weight of regularization, λ, goes to 0, the magnitudes of eigenvalues derived from the proposed scheme, denoted here by γ_1, γ_2, and γ_3, converge to the ones obtained from EDMD. Moreover, we know that as $\lambda \to +\infty$, the solution of (33) goes to 0. Accordingly, we expect that γ_1, γ_2, and γ_3 converge to 0. Based on Remarks 9 and 12, we expect to observe similar results for the case of $\mathcal{R}(\mathrm{K}) = \|\mathrm{K}\|_{\mathrm{F}}^2$ and $\mathcal{R}(\mathrm{K}) = \|\mathrm{K}\|_*$. Fig. 2 demonstrates these asymptotic phenomena. One can see that depending on the choice of regularization functions, we have different forms of convergence for γ_1, γ_2, and γ_3. Furthermore, comparing the cases $\mathcal{R}(\mathrm{K}) = \|\mathrm{K}\|^2$ and $\mathcal{R}(\mathrm{K}) = \|\mathrm{K}\|_{\mathrm{F}}^2$, when the regularization term is based on the Frobenius norm, we observe that as $\lambda \to \infty$, the convergence of $\gamma_1, \gamma_2, \gamma_3$ to 0 is with higher rate. Moreover, when nuclear norm is employed for the regularization term, γ_1, γ_2, and γ_3 converge to 0 faster than the two previous cases. These phenomena can be explained based on the inequality between the operator norm, Frobenius norm, and nuclear norm, i.e., $\|\mathrm{S}\| \leq \|\mathrm{S}\|_{\mathrm{F}} \leq \|\mathrm{S}\|_* \leq \infty$, which holds for all $\mathrm{S} \in \mathcal{L}(\mathcal{H}_{\mathbb{k}})$. Additionally, we can see that when $\mathcal{R}(\mathrm{K}) = \|\mathrm{K}\|_*$, the rank of operator $\hat{\mathrm{K}}$ drops as λ increases. The is due to the nature of nuclear norm, which leads to promoting low-rank solutions. △

Example 2: The Van der Pol oscillator is described as

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= \mu\left(1 - x_1^2\right) x_2 - x_1 + u \end{aligned} \tag{64}$$

where $\mu > 0$ is the damping coefficient and u is the input. To obtain a discrete-time system as in (1), we set $\mu = 0.5$ and $u = 0$ in (64), and employ forward Euler method with step size $\Delta t = 0.2\,\mathrm{s}$. We consider trajectories starting from initial points $\mathrm{x}_0^{(1)} = [-2,2]^\mathsf{T}$ and $\mathrm{x}_0^{(2)} = [0,-1]^\mathsf{T}$, respectively, with length $n_{\mathrm{s}}^{(1)} = 50$ and $n_{\mathrm{s}}^{(2)} = 50$ (see Fig. 3). We employ Matern kernel $\mathbb{k}$ with parameter $\nu_{\mathbb{k}} = \frac{5}{2}$ and length scale $\ell_{\mathbb{k}} = 1$ [50], and define observables $g_1, \ldots, g_{n_{\mathrm{g}}}$ as the sections of kernel $\mathbb{k}$ at points $\mathcal{P} = \{(0.5i, 0.5j) | i, j = -5, \ldots, 5\}$. To perform a Monte Carlo experiment, we corrupt the trajectories data points with zero mean white Gaussian additive noise. More precisely, we consider different values for noise variance to include low, medium, and high signal-to-noise ratio (SNR) levels, namely with 10, 20, and 30 dB. With respect to each of these SNR levels, 120 sequences of noise realization are generated for being added to the trajectories. Given observables data, we can estimate the Koopman operator using the EDMD approach and the regularized learning methods mentioned in the previous sections by solving their equivalent finite-dimensional optimization problems. To compare the performance of these methods, we employ test observables $\overline{g}_1, \ldots, \overline{g}_{n_{\overline{\mathrm{g}}}}$ specified as the sections of kernel $\mathbb{k}$ at points $\overline{\mathcal{P}} = \{(0.1i, 0.1j) | i, j = -5, \ldots, 5\}$, and the mean squared error for the estimated Koopman operator $\hat{\mathrm{K}}$ defined as

$$\mathrm{MSE}(\hat{\mathrm{K}}) = \frac{1}{n_{\overline{\mathrm{g}}}} \sum_{l=1}^{n_{\overline{\mathrm{g}}}} \int_R \Big((\hat{\mathrm{K}} \circ g_l)(\mathrm{x}) - g_l(F(\mathrm{x}))\Big)^2 \,\mathrm{dx} \tag{65}$$

where F is the vector field resulting from time discretization of Van der Pol dynamics (64), $R = [-2.5, 2.5] \times [-2.5, 2.5]$, and the integral is calculated numerically using a grid with $\Delta \mathrm{x} = (0.025, 0.025)$. Fig. 3 illustrates and compares estimation performance outcomes. One can observe that the introduced learning schemes outperform the EDMD method, i.e., the incorporation of regularization terms and constraints results in a more accurate estimation of the Koopman operator. Furthermore, learning techniques with regularization terms defined based on operator and Frobenius norms exhibit superior estimation performances compared to those with nuclear norm regularization. The observed outperformance can be due to the fact that nuclear norm dominates operator and Frobenius norms, and consequently, they have more effective impacts. The same arguments hold when the impact of rank constraint is compared with nuclear norm regularization. Moreover, we observe that the estimation performances generally improve as the SNR level increases, which is an expected phenomenon. The minor exception is the case of including rank constraint, which is possibly due to its nonconvexity. Finally, we can see that nuclear norm regularization and rank constraint are not as effective as the regularization terms specified by operator and Frobenius norm,

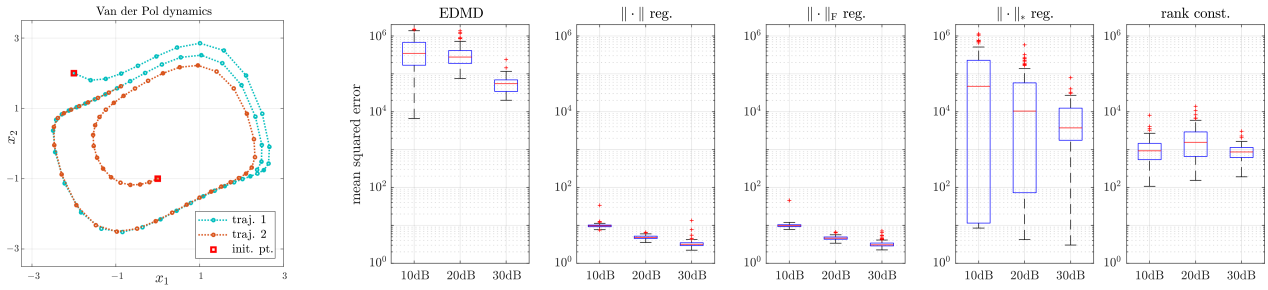

Fig. 3. Left: two trajectories of Van der Pol dynamics (64). Right: the box-plots for the mean squared error of different Koopman learning methods.

indicating that the original Koopman operator is probably not low-rank here. △

Example 3: The incorporation of potentially available side-information about the Koopman operator can improve the learning accuracy by shrinking the hypothesis space and rejecting spurious solution candidates. To demonstrate this feature, we employ a scaled version of the Nicholson–Bailey model for host-parasitoid dynamics [51] described as

$$\begin{aligned} x_{1,k+1} &= R_0\, x_{1,k}\,(1+2x_{2,k})^{-\frac{1}{2}} \\ x_{2,k+1} &= c\, x_{1,k}\left(1-(1+2x_{2,k})^{-\frac{1}{2}}\right) \end{aligned} \tag{66}$$

where $R_0 = 1.1$ and $c = 3$. We consider the trajectory starting from the initial point $\mathrm{x}_0 = [0.5, 0.05]^{\mathsf{T}}$ and with length $n_{\mathrm{s}} = 100$ (see Fig. 4). We employ Gaussian kernel (63) with $\ell_{\Bbbk} = 0.1$ [50]. The observables $g_1, \ldots, g_{n_{\mathrm{g}}}$ are defined as the sections of kernel $\Bbbk$ at points $\mathcal{P} = \{(0.3, 0.05) + (0.025i, 0.025j) | i, j = 0, \ldots, 8\}$. According to the behavior of the system shown in Fig. 4, one can conclude that the system is stable. Meanwhile, as we estimate the Koopman operator using the observables data through the EDMD approach and also the learning method with Frobenius norm regularization (44) where $\lambda = 10^{-6}$, we observe that the magnitude of dominant eigenvalues of the estimated Koopman operators is less than one. This observation confirms the discussed stability side-information. Following this, we perform a Monte Carlo experiment by corrupting the trajectory data with realizations of zero mean white Gaussian additive noise. The variance of noise is chosen such that the resulting SNR is 30 dB. We generate 450 sequence of noise realizations for being added to the trajectory data, and subsequently, the observables are evaluated on the noisy trajectory. Given the observables data, we repeat previous Koopman operator estimation schemes. Furthermore, to integrate the stability side-information, we modify the learning method with Frobenius norm regularization by including the stability-inducing constraint $\|\mathrm{K}\| \leq 1 - \varepsilon$ where $\varepsilon = 10^{-5}$. Fig. 4 (bottom) demonstrates the dominant eigenvalues of the estimated Koopman operators. We can see from Fig. 4 that when the stability-inducing constraint is used, the dominating eigenvalues are inside the unit circle, whereas they are mainly located outside the unit circle when the other techniques are implemented. Comparing the EDMD results, it can be seen that the dominant eigenvalues have a smaller magnitude when Frobenius norm regularization is used, which is potentially due to the inequality $\|\mathrm{S}\| \leq \|\mathrm{S}\|_{\mathrm{F}}$. Indeed, the Frobenius norm regularization partially incorporates the stability side-information.

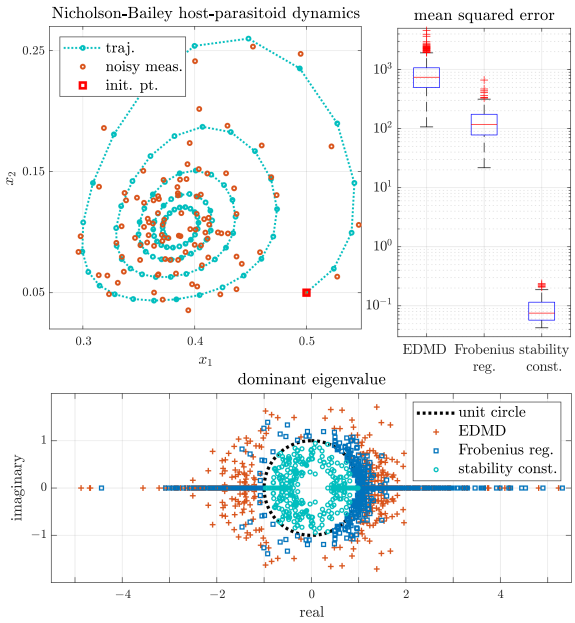

Fig. 4. Top-left: a trajectory of dynamics (66) and its noisy measurement with 30 dB SNR. Top-right: performance comparison for Koopman operator estimate using the EDMD approach, the learning scheme with Frobenius norm regularization, and its modified version with the stability-inducing constraint. Bottom: the dominant eigenvalues of estimated Koopman operators.

Similar to the previous example, we quantitatively compare the Koopman operator estimation results using test observables $\overline{g}_1, \ldots, \overline{g}_{n_{\overline{\mathrm{g}}}}$ defined as the sections of kernel $\Bbbk$ at points $\overline{\mathcal{P}} = \{(0.3, 0.05) + (0.01i, 0.01j) | i, j = 0, \ldots, 20\}$ and the mean squared error (65) calculated on region $R = [0.3, 0.5] \times [0.05, 0.25]$. Fig. 4 (top-right) compares the performance of discussed Koopman operator estimation schemes. One can see that the inclusion of stability side-information improves learning and estimation accuracy. △

Example 4: Consider the following convection-diffusion PDE:

$$\frac{\partial u}{\partial t}(\xi, t) = a\frac{\partial u}{\partial \xi}(\xi, t) + b\frac{\partial^2 u}{\partial \xi^2}(\xi, t) \tag{67}$$

where $(\xi, t) \in [0, 1] \times [0, \infty)$, $a = 1$, and $b = 0.1$. We discretize domains of ξ and t, respectively, with $\Delta\xi = 10^{-2}$ and $\Delta t = 10^{-4}$ to obtain discrete-time dynamics $\mathrm{x}^+ = F(\mathrm{x})$ with state

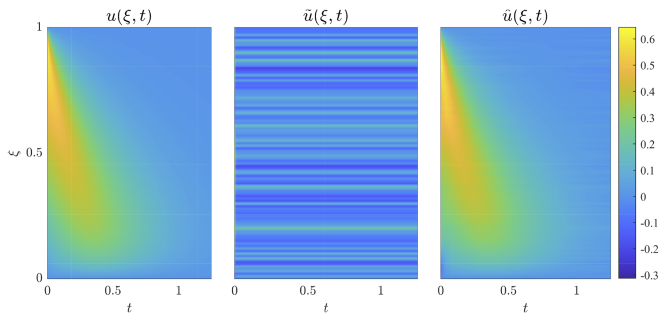

Fig. 5. Exact solution (left) of PDE (67) with initial condition $u(\xi,0) = 1-\mathrm{e}^{-\xi}$, and the approximate solutions derived from Koopman estimations, one obtained by the EDMD approach (middle) and the other via the learning method with Frobenius norm regularization (right).

dimension $n_{\mathrm{x}} = 101$. From the definition of Koopman operator, we know that $g(\mathrm{x}_n) = \mathrm{K}^n(\mathrm{x}_0)$, for each $n \in \mathbb{N}$. Thus, when x can be derived using the value of observables at x, one can employ the estimation of Koopman operator to obtain the solution of system for any arbitrarily initial condition. The precision of this solution indicates the accuracy of estimated Koopman operator. Accordingly, we consider the solution of discrete-time system that corresponds to the initial condition $u(\xi,0) = \sin(\pi\xi)$. Also, we employ kernel $\Bbbk(\mathrm{x},\mathrm{y}) = 1+\mathrm{x}^{\mathsf{T}}\mathrm{y}$, and specify $n_{\mathrm{g}} = 2n_{\mathrm{x}}$ observables $g_1,\ldots,g_{n_{\mathrm{g}}}$ as the sections of kernel $\Bbbk$ at points $\mathcal{P} = \{\mathrm{p}_1,\ldots,\mathrm{p}_{n_{\mathrm{g}}}\}$, which are randomly chosen from the standard normal distribution in $\mathbb{R}^{n_{\mathrm{x}}}$. We estimate the Koopman operator using data of the observables, and subsequently, we obtain the solution corresponding to the initial condition $u(\xi,0) = 1-\mathrm{e}^{-\xi}$ following the abovementioned discussion. To estimate the Koopman operator, we employ the EDMD approach and the learning method with Frobenius norm regularization, and denote the resulting approximated solutions respectively by $\tilde{u}$ and $\hat{u}$. To quantitatively compare these approximate solutions, we define solution mismatches as $\Delta\tilde{u} := \tilde{u}-u$ and $\Delta\hat{u} := \hat{u}-u$, and subsequently, their $\mathscr{L}^2$-norm is calculated on the region $[0,1]\times[0,1.25]$. The resulting values are $\|\Delta\tilde{u}\|_2 = 2.543\times10^{-1}$ and $\|\Delta\hat{u}\|_2 = 1.004\times10^{-2}$. In Fig. 5 , the approximate solutions are compared to the exact solution u. Fig. 5 shows that $\hat{u}$ is nearly identical to the exact solution, whereas $\tilde{u}$ appears to be considerably different. The quantitative evaluation and graphical comparison indicate that the Koopman operator estimation obtained by the learning method with Frobenius norm regularization is significantly more accurate compared to the one derived from the EDMD approach. △

X. Conclusion

In this article, we investigated the problem of learning Koopman operators for discrete-time autonomous systems. The learning problem was formulated as a constrained regularized optimization over the infinite-dimensional space of linear operators. We showed that a representer theorem holds for the Koopman learning problem, which allows a finite-dimension problem reformulation without any approximation and precision loss. Moreover, we investigated the incorporation of various forms of regularization and constraint in the Koopman operator learning problem, including the operator norm, the Frobenius norm, and rank. For each of these cases, we derived the corresponding finite-dimensional problem. We have also elaborated on the connection between these learning problems and the EDMD method. Finally, we have demonstrated the impact of regularizations, constraints, and side-information through several numerical examples.

Appendix

A. Proof of Theorem 3

The existence and uniqueness of the solution $\hat{\mathrm{K}}_{\mathcal{W}}$ follows from the same lines of arguments as in the proof of Theorem 1. Define the linear subspace $\widetilde{\mathcal{W}}$ as $\widetilde{\mathcal{W}} := \operatorname{span}\{\Pi_{\mathcal{W}}v_1,\ldots,\Pi_{\mathcal{W}}v_{n_{\mathrm{s}}}\}$, and, let $\Pi_{\widetilde{\mathcal{W}}}$ be the projection operator on $\widetilde{\mathcal{W}}$. For $k\in[n_{\mathrm{s}}]$, we know that $\Pi_{\mathcal{W}}v_k \in \mathcal{W}$, and therefore, $\widetilde{\mathcal{W}}$ is a subspace of $\mathcal{W}$. Define operator S as $\mathrm{S} := \Pi_{\widetilde{\mathcal{W}}}\hat{\mathrm{K}}_{\mathcal{W}}\Pi_{\mathcal{G}}$. Since $\widetilde{\mathcal{W}}\subseteq\mathcal{W}$, we have $\mathrm{S}\in\mathcal{L}_{\mathcal{W}}$. From the definition of $\Pi_{\mathcal{G}}$, we know that $\Pi_{\mathcal{G}}g_l = g_l$, for $l\in[n_{\mathrm{g}}]$. Thus, for any k and l, we have

$$\begin{aligned}\langle v_k,\mathrm{S}g_l\rangle &= \left\langle v_k,\Pi_{\widetilde{\mathcal{W}}}\hat{\mathrm{K}}_{\mathcal{W}}\Pi_{\mathcal{G}}g_l\right\rangle\\ &= \left\langle v_k,\Pi_{\widetilde{\mathcal{W}}}\hat{\mathrm{K}}_{\mathcal{W}}g_l\right\rangle = \left\langle\Pi^*_{\widetilde{\mathcal{W}}}v_k,\hat{\mathrm{K}}_{\mathcal{W}}g_l\right\rangle\end{aligned} \tag{68}$$

where $\Pi^*_{\widetilde{\mathcal{W}}}$ is the adjoint of $\Pi_{\widetilde{\mathcal{W}}}$. Since $\widetilde{\mathcal{W}}$ is a finite dimensional subspace, it is closed and the projection operator $\Pi_{\widetilde{\mathcal{W}}}$ is self-adjoint, i.e., $\Pi^*_{\widetilde{\mathcal{W}}} = \Pi_{\widetilde{\mathcal{W}}}$. Also, from $\widetilde{\mathcal{W}}\subseteq\mathcal{W}$, we know that $\mathcal{W}^{\perp}\subseteq\widetilde{\mathcal{W}}^{\perp}$, and subsequently, we have $\Pi_{\widetilde{\mathcal{W}}}\Pi_{\mathcal{W}^{\perp}} = 0$. Accordingly, one can see that $\Pi_{\widetilde{\mathcal{W}}} - \Pi_{\widetilde{\mathcal{W}}}\Pi_{\mathcal{W}} = \Pi_{\widetilde{\mathcal{W}}}(\mathbb{I}-\Pi_{\mathcal{W}}) = \Pi_{\widetilde{\mathcal{W}}}\Pi_{\mathcal{W}^{\perp}} = 0$. Thus, for each $k\in[n_{\mathrm{s}}]$, we have

$$\Pi^*_{\widetilde{\mathcal{W}}}v_k = \Pi_{\widetilde{\mathcal{W}}}v_k = \Pi_{\widetilde{\mathcal{W}}}\Pi_{\mathcal{W}}v_k = \Pi_{\mathcal{W}}v_k \tag{69}$$

where the last equality is due to $\Pi_{\mathcal{W}}v_k\in\widetilde{\mathcal{W}}$. Note that $\mathcal{W}$ is a closed subspace, and subsequently, $\Pi_{\mathcal{W}}$ is a self-adjoint operator, i.e., $\Pi^*_{\mathcal{W}} = \Pi_{\mathcal{W}}$. Accordingly, from (68) and (69), we can see that

$$\begin{aligned}\langle v_k,\mathrm{S}g_l\rangle &= \left\langle\Pi_{\mathcal{W}}v_k,\hat{\mathrm{K}}_{\mathcal{W}}g_l\right\rangle\\ &= \left\langle v_k,\Pi^*_{\mathcal{W}}\hat{\mathrm{K}}_{\mathcal{W}}g_l\right\rangle = \left\langle v_k,\Pi_{\mathcal{W}}\hat{\mathrm{K}}_{\mathcal{W}}g_l\right\rangle\end{aligned} \tag{70}$$

for each k and l. Due to the definition of $\mathcal{L}_{\mathcal{W}}$ in (20) and since $\hat{\mathrm{K}}_{\mathcal{W}}\in\mathcal{L}_{\mathcal{W}}$, we know that $\hat{\mathrm{K}}_{\mathcal{W}}g_l\in\mathcal{W}$, and subsequently, $\Pi_{\mathcal{W}}\hat{\mathrm{K}}_{\mathcal{W}}g_l = \hat{\mathrm{K}}_{\mathcal{W}}g_l$, for $l\in[n_{\mathrm{g}}]$. Hence, from (70), we have

$$\begin{aligned}\mathcal{E}(\mathrm{S}) &= \sum_{k=1}^{n_{\mathrm{s}}}\sum_{l=1}^{n_{\mathrm{g}}}(y_{kl}-\langle v_k,\mathrm{S}g_l\rangle)^2\\ &= \sum_{k=1}^{n_{\mathrm{s}}}\sum_{l=1}^{n_{\mathrm{g}}}(y_{kl}-\langle v_k,\hat{\mathrm{K}}_{\mathcal{W}}g_l\rangle)^2 = \mathcal{E}(\hat{\mathrm{K}}_{\mathcal{W}}).\end{aligned}$$

Similar to the proof of Theorem 1, one can show that $\|\mathrm{S}\|^2\le\|\hat{\mathrm{K}}_{\mathcal{W}}\|^2$, and subsequently, one can see that $\mathcal{E}(\mathrm{S})+\lambda\|\mathrm{S}\|^2\le\mathcal{E}(\hat{\mathrm{K}}_{\mathcal{W}})+\lambda\|\hat{\mathrm{K}}_{\mathcal{W}}\|^2$. From the uniqueness of the solution of (25), we have $\hat{\mathrm{K}}_{\mathcal{W}} = \mathrm{S} = \Pi_{\mathcal{V}}\hat{\mathrm{K}}_{\mathcal{W}}\Pi_{\mathcal{G}}$. Due to the linearity of operator $\hat{\mathrm{K}}_{\mathcal{W}}$, it follows that there exist $a_{kl}\in\mathbb{R}$, for $k\in[n_{\mathrm{s}}]$ and $l\in[n_{\mathrm{g}}]$, such that $\hat{\mathrm{K}}_{\mathcal{W}} = \sum_{k=1}^{n_{\mathrm{s}}}\sum_{l=1}^{n_{\mathrm{g}}}a_{kl}(\Pi_{\mathcal{W}}v_k)\otimes g_l$, i.e., we have (26). Considering $\hat{\mathrm{K}}_{\mathcal{W}}$ in this parametric form and due to

the linearity of inner product, for each $i \in [n_s]$ and $j \in [n_g]$, it follows that

$$\begin{aligned}(\hat{\mathrm{K}} g_j)(\mathrm{x}_{i-1}) &= \sum_{k=1}^{n_s}\sum_{l=1}^{n_g} a_{kl}\,\langle v_i, ((\Pi_{\mathcal{W}} v_k)\otimes g_l)\, g_j\rangle \\ &= \sum_{k=1}^{n_s}\sum_{l=1}^{n_g} a_{kl}\,\langle v_i, \Pi_{\mathcal{W}} v_k\rangle\langle g_l, g_j\rangle\rangle \\ &= \sum_{k=1}^{n_s}\sum_{l=1}^{n_g} \langle \Pi_{\mathcal{W}} v_i, \Pi_{\mathcal{W}} v_k\rangle\, a_{kl}\,\langle g_l, g_j\rangle\end{aligned}$$

where the last equality is due to the fact that $\langle u, \Pi_{\mathcal{W}} v\rangle = \langle \Pi_{\mathcal{W}} u, \Pi_{\mathcal{W}} v\rangle$, for any $v, u \in \mathcal{H}$. Accordingly, we have $[(\hat{\mathrm{K}} g_j)(\mathrm{x}_{i-1})]_{i=1,j=1}^{n_s,\,n_g} = \mathrm{W}_{\mathcal{V}}\mathrm{AG}$. Thus, following the calculation steps similar to those in the proof of Theorem 1, we can show that A can be obtained by solving convex program (27).■

B. Proof of Theorem 4

For each $k \in [n_s]$, there exist $q_{k1}, \ldots, q_{kn_w}$ such that

$$\Pi_{\mathcal{W}} v_k = \sum_{j=1}^{n_w} q_{kj} w_j = \operatorname*{argmin}_{w\in\mathcal{W}} \frac{1}{2}\|v_k - w\|^2. \tag{71}$$

Define matrix Q as $\mathrm{Q} = [q_{kj}]_{k=1,j=1}^{n_s,\,n_w}$, and let the row vectors q_k and p_k be, respectively, defined as $\mathrm{q}_k = [q_{kj}]_{j=1}^{n_w}$ and $\mathrm{p}_k = [\langle v_k, w_j\rangle]_{j=1}^{n_w}$. From (71), we can easily see that $\mathrm{q}_k = \mathrm{p}_k \mathrm{W}^{-1}$. Also, due to the definition of vectors $v_1, \ldots, v_{n_s}$, we know that $\mathrm{P}_{\mathcal{W}} = [p_1^{\mathsf{T}}, \ldots, p_{n_s}^{\mathsf{T}}]^{\mathsf{T}}$, which implies that $\mathrm{Q} = \mathrm{P}_{\mathcal{W}}\mathrm{W}^{-1}$. From (71), which says that $\Pi_{\mathcal{W}} v_k = \sum_{j=1}^{n_w} q_{kj} w_j$, one can see that $\hat{\mathrm{K}}_{\mathcal{W}}$ in (26) has a representation as in (28). Moreover, we have

$$\begin{aligned}\hat{\mathrm{K}}_{\mathcal{W}} &= \sum_{j=1}^{n_w}\sum_{l=1}^{n_g}\left(\sum_{k=1}^{n_s} q_{kj} a_{kl}\right) w_j \otimes g_l \\ &= \sum_{j=1}^{n_w}\sum_{l=1}^{n_g} [\mathrm{Q}^{\mathsf{T}}\mathrm{A}]_{(j,l)}\, w_j \otimes g_l\end{aligned}$$

which yields $\mathrm{C} = \mathrm{Q}^{\mathsf{T}}\mathrm{A}$. Also, for each $k, i \in [n_s]$, one has

$$\begin{aligned}[\mathrm{W}_{\mathcal{V}}]_{(k,i)} &= \langle \Pi_{\mathcal{W}} v_k, \Pi_{\mathcal{W}} v_i\rangle = \left\langle \textstyle\sum_{j=1}^{n_w} q_{kj} w_j, \sum_{l=1}^{n_w} q_{il} w_l \right\rangle \\ &= \textstyle\sum_{j=1}^{n_w}\sum_{l=1}^{n_w} q_{kj}\langle w_j, w_l\rangle q_{il} = [\mathrm{QWQ}^{\mathsf{T}}]_{(k,i)}\end{aligned} \tag{72}$$

and, consequently, we have $\mathrm{W}_{\mathcal{V}} = \mathrm{QWQ}^{\mathsf{T}}$. We, thus, get $\mathrm{W}_{\mathcal{V}} = \mathrm{P}_{\mathcal{W}}\mathrm{W}^{-1}\mathrm{P}_{\mathcal{W}}^{\mathsf{T}}$. Subsequently, it follows that

$$\mathrm{P}_{\mathcal{W}}\mathrm{C} = \mathrm{P}_{\mathcal{W}}\mathrm{Q}^{\mathsf{T}}\mathrm{A} = \mathrm{P}_{\mathcal{W}}\mathrm{W}^{-1}\mathrm{P}_{\mathcal{W}}^{\mathsf{T}}\mathrm{A} = \mathrm{W}_{\mathcal{V}}\mathrm{A} \tag{73}$$

which gives $\|\mathrm{P}_{\mathcal{W}}\mathrm{CG} - \mathrm{Y}\|_{\mathrm{F}}^2 = \|\mathrm{W}_{\mathcal{V}}\mathrm{AG} - \mathrm{Y}\|_{\mathrm{F}}^2$. Furthermore, from $\mathrm{A}^{\mathsf{T}}\mathrm{W}_{\mathcal{V}}\mathrm{A} = \mathrm{A}^{\mathsf{T}}\mathrm{QWQ}^{\mathsf{T}}\mathrm{A} = \mathrm{C}^{\mathsf{T}}\mathrm{WC}$, we know that

$$\begin{aligned}\|\mathrm{W}^{\frac{1}{2}}\mathrm{CG}^{\frac{1}{2}}\|^2 &= \sup_{\mathrm{z}\in\mathbb{R}^{n_g},\|\mathrm{z}\|\le 1} \mathrm{z}^{\mathsf{T}}\mathrm{G}^{\frac{1}{2}}\mathrm{C}^{\mathsf{T}}\mathrm{WCG}^{\frac{1}{2}}\mathrm{z} \\ &= \sup_{\mathrm{z}\in\mathbb{R}^{n_g},\|\mathrm{z}\|\le 1} \mathrm{z}^{\mathsf{T}}\mathrm{G}^{\frac{1}{2}}\mathrm{A}^{\mathsf{T}}\mathrm{W}_{\mathcal{V}}\mathrm{AG}^{\frac{1}{2}}\mathrm{z} = \|\mathrm{W}_{\mathcal{V}}^{\frac{1}{2}}\mathrm{AG}^{\frac{1}{2}}\|^2.\end{aligned}$$

Thus, using the mentioned change-of-variable in (27), we get the convex program (29). ■

C. Proof of Theorem 10

We know that $\mathrm{K} = 0$ is a feasible point for (44). Therefore, for the optimal solution of (44), we need to have

$$\lambda\|\hat{\mathrm{K}}\|_{\mathrm{F}}^2 \le \mathcal{E}(\hat{\mathrm{K}}) + \lambda\|\hat{\mathrm{K}}\|_{\mathrm{F}}^2 \le \mathcal{E}(0) + \lambda\|0\|_{\mathrm{F}}^2 = \|\mathrm{Y}\|_{\mathrm{F}}^2. \tag{74}$$

Accordingly, (44) is equivalent to the following problem:

$$\begin{aligned}\min_{\mathrm{K}\in\mathcal{F}}\quad & \mathcal{E}(\mathrm{K}) + \lambda\|\mathrm{K}\|_{\mathrm{F}}^2 \\ \text{s.t.}\quad & \mathrm{K}\in\mathcal{C}\end{aligned} \tag{75}$$

where $\mathcal{C}$ is defined as $\mathcal{C} := \{\mathrm{S}\in\mathcal{L}(\mathcal{H}) \mid \|\mathrm{S}\|_{\mathrm{F}} \le \frac{1}{\sqrt{\lambda}}\|\mathrm{Y}\|_{\mathrm{F}}\}$. Let $\mathrm{S}\in\mathcal{L}(\mathcal{H})$, and $\{b_k\}_{k=1}^{\infty}$ be an orthonormal basis for $\mathcal{H}$ such that $\mathcal{G} = \operatorname{span}\{b_1, \ldots, b_{\overline{n}_g}\}$, where $\overline{n}_g \le n_g$. If $k \le \overline{n}_g$, we have $\|\Pi_{\mathcal{Z}}\mathrm{S}\Pi_{\mathcal{G}} b_k\| = \|\Pi_{\mathcal{Z}}\mathrm{S}b_k\|$, and since $\|\Pi_{\mathcal{Z}}\| \le 1$, it follows that $\|\Pi_{\mathcal{Z}}\mathrm{S}\Pi_{\mathcal{G}} b_k\| \le \|\mathrm{S}b_k\|$. If $k > \overline{n}_g$, then $\|\Pi_{\mathcal{Z}}\mathrm{S}\Pi_{\mathcal{G}} b_k\| = 0$. Therefore, due to the definition of Frobenius norm, we have

$$\|\Pi_{\mathcal{Z}}\mathrm{S}\Pi_{\mathcal{G}}\|_{\mathrm{F}}^2 = \sum_{k=1}^{\infty}\|\Pi_{\mathcal{Z}}\mathrm{S}\Pi_{\mathcal{G}} b_k\|^2 \le \sum_{k=1}^{\infty}\|\mathrm{S}b_k\|^2 = \|\mathrm{S}\|_{\mathrm{F}}^2.$$

Thus, for any $\mathrm{S}\in\mathcal{L}(\mathcal{H})$, we have $\mathcal{R}(\Pi_{\mathcal{Z}}\mathrm{S}\Pi_{\mathcal{G}}) = \|\Pi_{\mathcal{Z}}\mathrm{S}\Pi_{\mathcal{G}}\|_{\mathrm{F}}^2 \le \|\mathrm{S}\|_{\mathrm{F}}^2 = \mathcal{R}(\mathrm{S})$. Moreover, for each $\mathrm{S}\in\mathcal{C}$, one can see that $\|\Pi_{\mathcal{Z}}\mathrm{S}\Pi_{\mathcal{G}}\|_{\mathrm{F}} \le \|\mathrm{S}\|_{\mathrm{F}} \le \frac{1}{\sqrt{\lambda}}\|\mathrm{Y}\|_{\mathrm{F}}$, which implies that $\Pi_{\mathcal{Z}}\mathrm{S}\Pi_{\mathcal{G}} \in \mathcal{C}$. Therefore, $\delta_{\mathcal{C}}(\Pi_{\mathcal{Z}}\mathrm{S}\Pi_{\mathcal{G}}) \le \delta_{\mathcal{C}}(\mathrm{S})$, and hence, Assumption 2 holds. By definition, we know that $\mathcal{C}\subset\operatorname{dom}(\mathcal{R})$ and $0\in\mathcal{D} := \mathcal{F}\cap\operatorname{dom}(\mathcal{R})\cap\mathcal{C} = \mathcal{F}\cap\mathcal{C}$. Since $\|\cdot\|_{\mathrm{F}}$ is a norm on $\operatorname{dom}(\mathcal{R})$ and $\mathcal{C}\subset\operatorname{dom}(\mathcal{R})$, it follows that $\mathcal{C}$ is a convex set. Let $\{\mathrm{S}_n\}_{n\in\mathbb{N}}\subset\mathcal{L}(\mathcal{H})$ be a sequence such that $\lim_{n\to\infty}\mathrm{S}_n = \mathrm{S}\in\mathcal{L}(\mathcal{H})$ in norm topology. For any $k\in\mathbb{N}$, we know that $\lim_{n\to\infty}\|\mathrm{S}_n b_k\|^2 = \|\mathrm{S}b_k\|^2$. From Fatou's lemma, it follows that

$$\begin{aligned}\|\mathrm{S}\|_{\mathrm{F}}^2 &= \textstyle\sum_{k=1}^{\infty}\|\mathrm{S}b_k\|^2 = \sum_{k=1}^{\infty}\liminf_n\|\mathrm{S}_n b_k\|^2 \\ &\le \liminf_n\textstyle\sum_{k=1}^{\infty}\|\mathrm{S}_n b_k\|^2 = \liminf_n\|\mathrm{S}_n\|_{\mathrm{F}}^2.\end{aligned} \tag{76}$$

Therefore, we have $\mathcal{R}(\mathrm{S}) \le \liminf_n \mathcal{R}(\mathrm{S}_n)$, and $\mathcal{R}$ is lower semicontinuous. Moreover, if $\{\mathrm{S}_n\}_{n\in\mathbb{N}}\subset\mathcal{C}$, then $\|\mathrm{S}_n\|_{\mathrm{F}} \le \frac{1}{\sqrt{\lambda}}\|\mathrm{Y}\|_{\mathrm{F}}$, for each n. Subsequently, due to (76), we have $\|\mathrm{S}\|_{\mathrm{F}} \le \frac{1}{\sqrt{\lambda}}\|\mathrm{Y}\|_{\mathrm{F}}$, which implies that $\mathrm{S}\in\mathcal{C}$. Hence, $\mathcal{C}$ and $\mathcal{D} = \mathcal{F}\cap\mathcal{C}$ are nonempty, closed, and convex sets. Due to $\|\mathrm{S}\| \le \|\mathrm{S}\|_{\mathrm{F}}$, we know that $\mathcal{R}$ is coercive, which implies that $\overline{\mathcal{R}}$ is coercive as well. Furthermore, it follows from Lemma 16 that $\mathcal{R}$ is strictly convex. Therefore, due to Theorem 8, (75) admits a unique solution with the parametric form in (40). This solution coincides with the unique solution of (44) due to the equivalency of the corresponding programs. Since for any $h_1, h_2\in\mathcal{H}$, we have $(h_1\otimes h_2)^* = h_2\otimes h_1$, for $\hat{\mathrm{K}}$ in the given parametric form, we know that $\hat{\mathrm{K}}^* = \sum_{k=1}^{n_z}\sum_{l=1}^{n_g} a_{kl}\, g_l\otimes z_k$. One can easily see that

$$(h_1\otimes h_2)(h_3\otimes h_4) = \langle h_2, h_3\rangle(h_1\otimes h_4) \tag{77}$$

for any $h_1, h_2, h_3, h_4\in\mathcal{H}$. Accordingly, we have

$$\begin{aligned}\hat{\mathrm{K}}^*\hat{\mathrm{K}} &= \textstyle\sum_{k=1}^{n_z}\sum_{j=1}^{n_g}\sum_{l=1}^{n_g}\sum_{i=1}^{n_z} a_{kl}a_{ij}(g_l\otimes z_k)(z_i\otimes g_j) \\ &= \textstyle\sum_{j=1}^{n_g}\sum_{l=1}^{n_g}(g_l\otimes g_j)\sum_{k=1}^{n_z}\sum_{i=1}^{n_z} a_{kl}\langle z_k, z_i\rangle a_{ij} \\ &= \textstyle\sum_{j=1}^{n_g}\sum_{l=1}^{n_g}(g_l\otimes g_j)[\mathrm{A}^{\mathsf{T}}\mathrm{ZA}]_{(l,j)}.\end{aligned} \tag{78}$$

Since $\operatorname{tr}(g_l\otimes g_j) = \langle g_l, g_j\rangle$, for each $j, l\in[n_g]$, it follows from (78) and $\|\hat{\mathrm{K}}\|_{\mathrm{F}}^2 = \operatorname{tr}(\hat{\mathrm{K}}^*\hat{\mathrm{K}})$ that

$$\begin{aligned}\|\hat{\mathrm{K}}\|_{\mathrm{F}}^2 &= \textstyle\sum_{j=1}^{n_g}\sum_{l=1}^{n_g}\langle g_l, g_j\rangle[\mathrm{A}^{\mathsf{T}}\mathrm{ZA}]_{(l,j)} \\ &= \operatorname{tr}(\mathrm{GA}^{\mathsf{T}}\mathrm{ZA}) = \operatorname{tr}(\mathrm{G}^{\frac{1}{2}}\mathrm{A}^{\mathsf{T}}\mathrm{Z}^{\frac{1}{2}}\mathrm{Z}^{\frac{1}{2}}\mathrm{AG}^{\frac{1}{2}}) = \|\mathrm{Z}^{\frac{1}{2}}\mathrm{AG}^{\frac{1}{2}}\|_{\mathrm{F}}^2.\end{aligned} \tag{79}$$

Using calculation steps similar to those in the proof of Theorem 1, one can see $\mathcal{E}(\hat{\mathrm{K}}) = \|\mathrm{ZAG} - \mathrm{Y}\|_{\mathrm{F}}^2$. Replacing these terms, optimization (47) results. ■

D. Proof of Theorem 11

We know that (51) is a special case of (38) where $\mathcal{R}\equiv 0$, and $\mathcal{C}$ is given in (50). Accordingly, we have $\overline{\mathcal{R}} = \delta_{\mathcal{C}}$. For any

operator $\mathrm{S} \in \mathcal{H}$, we know that

$$\begin{aligned}\operatorname{rank}(\Pi_{\mathcal{Z}}\mathrm{S}\Pi_{\mathcal{G}}) &\leq \min\{\operatorname{rank}(\Pi_{\mathcal{Z}}), \operatorname{rank}(\mathrm{S}), \operatorname{rank}(\Pi_{\mathcal{G}})\}\\ &\leq \operatorname{rank}(\mathrm{S}).\end{aligned}$$

Hence, $\mathrm{S} \in \mathcal{C}$ implies that $\Pi_{\mathcal{Z}}\mathrm{S}\Pi_{\mathcal{G}} \in \mathcal{C}$, and, we have

$$\overline{\mathcal{R}}(\Pi_{\mathcal{Z}}\mathrm{S}\Pi_{\mathcal{G}}) = \delta_{\mathcal{C}}(\Pi_{\mathcal{Z}}\mathrm{S}\Pi_{\mathcal{G}}) \leq \delta_{\mathcal{C}}(\mathrm{S}) = \overline{\mathcal{R}}(\mathrm{S}). \tag{80}$$

Therefore, the Assumption 2 holds for (51). Accordingly, due to Theorem 8, if (5) admits a solution, it has also a solution $\hat{\mathrm{K}}$ with parametric form given in (40). By taking orthonormal bases for $\mathcal{Z}$ and $\mathcal{G}$, one can easily show that $\operatorname{rank}(\hat{\mathrm{K}}) = \operatorname{rank}(\mathrm{ZAG})$. More precisely, let $\{b_1, \ldots, b_{\overline{n}_{\mathrm{g}}}\}$ and $\{p_1, \ldots, p_{\overline{n}_{\mathrm{z}}}\}$ be two sets of orthonormal vectors such that $\mathcal{G} = \operatorname{span}\{b_1, \ldots, b_{\overline{n}_{\mathrm{g}}}\}$ and $\mathcal{Z} = \operatorname{span}\{p_1, \ldots, p_{\overline{n}_{\mathrm{z}}}\}$. Then, for each $l \in [n_{\mathrm{g}}]$ and $k \in [n_{\mathrm{z}}]$, there exist real numbers $e_{l1}, \ldots, e_{l\overline{n}_{\mathrm{g}}}$ and $q_{k1}, \ldots, q_{k\overline{n}_{\mathrm{z}}}$ such that $g_l = \sum_{j=1}^{\overline{n}_{\mathrm{g}}} e_{lj} b_j$ and $z_k = \sum_{i=1}^{\overline{n}_{\mathrm{g}}} q_{ki} p_i$. Accordingly, one can see that

$$\begin{aligned}\hat{\mathrm{K}} &= \textstyle\sum_{i=1}^{\overline{n}_{\mathrm{z}}} \sum_{j=1}^{\overline{n}_{\mathrm{g}}} \left(\sum_{k=1}^{n_{\mathrm{z}}} \sum_{l=1}^{n_{\mathrm{g}}} q_{ki}\, a_{kl}\, e_{lj}\right) (p_i \otimes b_j)\\ &= \textstyle\sum_{i=1}^{\overline{n}_{\mathrm{z}}} \sum_{j=1}^{\overline{n}_{\mathrm{g}}} [\mathrm{Q}^{\mathsf{T}}\mathrm{AE}]_{(i,j)} (p_i \otimes b_j)\end{aligned} \tag{81}$$

where matrices E and Q are defined, respectively, as $\mathrm{E} = [e_{lj}]_{l=1,j=1}^{n_{\mathrm{g}},\overline{n}_{\mathrm{g}}}$ and $\mathrm{Q} = [q_{ki}]_{k=1,i=1}^{n_{\mathrm{z}},\overline{n}_{\mathrm{z}}}$. From (81), we know that $\operatorname{rank}(\hat{\mathrm{K}}) = \operatorname{rank}(\mathrm{Q}^{\mathsf{T}}\mathrm{AE})$. Also, we can obtain the Gramian matrices G and Z, respectively, as $\mathrm{G} = \mathrm{EE}^{\mathsf{T}}$ and $\mathrm{Z} = \mathrm{QQ}^{\mathsf{T}}$. Thus, from Lemma 18 in Appendix H, we have

$$\begin{aligned}\operatorname{rank}(\hat{\mathrm{K}}) &= \operatorname{rank}\left(\mathrm{Q}^{\mathsf{T}}\mathrm{AE}\right) = \operatorname{rank}\left(\mathrm{Q}^{\mathsf{T}}\mathrm{AEE}^{\mathsf{T}}\right)\\ &= \operatorname{rank}\left(\mathrm{QQ}^{\mathsf{T}}\mathrm{AEE}^{\mathsf{T}}\right) = \operatorname{rank}(\mathrm{ZAG}).\end{aligned} \tag{82}$$

Using calculation steps similar to those in the proof of Theorem 1, we have $\mathcal{E}(\hat{\mathrm{K}}) = \|\mathrm{ZAG} - \mathrm{Y}\|_{\mathrm{F}}^2$. Replacing these terms, we obtain optimization (52). ■

E. Proof of Theorem 12

For program (56), $\mathrm{K} = 0$ is a feasible solution. Therefore, for the optimal solution of (56), the following inequality holds:

$$\lambda\|\hat{\mathrm{K}}\|_* \leq \mathcal{E}(\hat{\mathrm{K}}) + \lambda\|\hat{\mathrm{K}}\|_* \leq \mathcal{E}(0) + \lambda\|0\|_* = \|\mathrm{Y}\|_{\mathrm{F}}^2. \tag{83}$$

By virtue of (83), we define $\mathcal{C}$ as $\mathcal{C} := \{\mathrm{S} \in \mathcal{L}(\mathcal{H}) \mid \|\mathrm{S}\|_* \leq \lambda^{-1}\|\mathrm{Y}\|_{\mathrm{F}}\}$, and, consider constrained optimization problem

$$\min_{\mathrm{K} \in \mathcal{F} \cap \mathcal{C}} \ \mathcal{E}(\mathrm{K}) + \lambda\|\mathrm{K}\|_* \tag{84}$$

which is equivalent to (56). Let $\mathrm{S} \in \mathcal{L}(\mathcal{H})$, and $\{b_k\}_{k=1}^{\infty}$ be an orthonormal basis for $\mathcal{H}$ such that $\mathcal{G} = \operatorname{span}\{b_1, \ldots, b_{\overline{n}_{\mathrm{g}}}\}$. Note that $\Pi_{\mathcal{G}} b_k = b_k$, if $k \leq \overline{n}_{\mathrm{g}}$, and, $\Pi_{\mathcal{G}} b_k = 0$, otherwise. Accordingly, from (55), (45), and $\Pi_{\mathcal{G}}^* = \Pi_{\mathcal{G}}$, we have

$$\begin{aligned}\|\Pi_{\mathcal{Z}}\mathrm{S}\Pi_{\mathcal{G}}\|_* &= \sup_{\|\mathrm{C}\|\leq 1, \mathrm{C}\in\mathcal{K}(\mathcal{H})} \left|\operatorname{tr}(\mathrm{C}\Pi_{\mathcal{Z}}\mathrm{S}\Pi_{\mathcal{G}})\right|\\ &= \sup_{\|\mathrm{C}\|\leq 1, \mathrm{C}\in\mathcal{K}(\mathcal{H})} \left|\textstyle\sum_{k=1}^{\infty} \langle \mathrm{C}\Pi_{\mathcal{Z}}\mathrm{S}\Pi_{\mathcal{G}} b_k, b_k\rangle\right|\\ &= \sup_{\|\mathrm{C}\|\leq 1, \mathrm{C}\in\mathcal{K}(\mathcal{H})} \left|\textstyle\sum_{k=1}^{\overline{n}_{\mathrm{g}}} \langle \mathrm{C}\Pi_{\mathcal{Z}}\mathrm{S}\Pi_{\mathcal{G}} b_k, b_k\rangle\right|\\ &= \sup_{\|\mathrm{C}\|\leq 1, \mathrm{C}\in\mathcal{K}(\mathcal{H})} \left|\textstyle\sum_{k=1}^{\overline{n}_{\mathrm{g}}} \langle \mathrm{C}\Pi_{\mathcal{Z}}\mathrm{S} b_k, b_k\rangle\right|\\ &= \sup_{\|\mathrm{C}\|\leq 1, \mathrm{C}\in\mathcal{K}(\mathcal{H})} \left|\textstyle\sum_{k=1}^{\overline{n}_{\mathrm{g}}} \langle \mathrm{C}\Pi_{\mathcal{Z}}\mathrm{S} b_k, \Pi_{\mathcal{G}} b_k\rangle\right|\\ &= \sup_{\|\mathrm{C}\|\leq 1, \mathrm{C}\in\mathcal{K}(\mathcal{H})} \left|\textstyle\sum_{k=1}^{\infty} \langle \mathrm{C}\Pi_{\mathcal{Z}}\mathrm{S} b_k, \Pi_{\mathcal{G}} b_k\rangle\right|\\ &= \sup_{\|\mathrm{C}\|\leq 1, \mathrm{C}\in\mathcal{K}(\mathcal{H})} \left|\textstyle\sum_{k=1}^{\infty} \langle \Pi_{\mathcal{G}}\mathrm{C}\Pi_{\mathcal{Z}}\mathrm{S} b_k, b_k\rangle\right|\\ &\leq \sup_{\|\mathrm{C}\|\leq 1, \mathrm{C}\in\mathcal{K}(\mathcal{H})} \left|\textstyle\sum_{k=1}^{\infty} \langle \mathrm{CS} b_k, b_k\rangle\right|\\ &= \sup_{\|\mathrm{C}\|\leq 1, \mathrm{C}\in\mathcal{K}(\mathcal{H})} \left|\operatorname{tr}(\mathrm{CS})\right| = \|\mathrm{S}\|_*\end{aligned}$$

where the inequality is due to the fact that $\Pi_{\mathcal{G}}\mathrm{C}\Pi_{\mathcal{Z}} \in \mathcal{K}(\mathcal{H})$ with $\|\Pi_{\mathcal{G}}\mathrm{C}\Pi_{\mathcal{Z}}\| \leq 1$, when $\mathrm{C} \in \mathcal{K}(\mathcal{H})$ with $\|\mathrm{C}\| \leq 1$. Therefore, for any $\mathrm{S} \in \mathcal{L}(\mathcal{H})$, one can see that $\mathcal{R}(\Pi_{\mathcal{Z}}\mathrm{S}\Pi_{\mathcal{G}}) = \|\Pi_{\mathcal{Z}}\mathrm{S}\Pi_{\mathcal{G}}\|_* \leq \|\mathrm{S}\|_* = \mathcal{R}(\mathrm{S})$. Moreover, for $\mathrm{S} \in \mathcal{C}$, we have $\|\Pi_{\mathcal{Z}}\mathrm{S}\Pi_{\mathcal{G}}\|_* \leq \|\mathrm{S}\|_* \leq \lambda^{-1}\|\mathrm{Y}\|_{\mathrm{F}}$, and, subsequently, it follows that $\Pi_{\mathcal{Z}}\mathrm{S}\Pi_{\mathcal{G}} \in \mathcal{C}$. Therefore, $\delta_{\mathcal{C}}(\Pi_{\mathcal{Z}}\mathrm{S}\Pi_{\mathcal{G}}) \leq \delta_{\mathcal{C}}(\mathrm{S})$, and hence, Assumption 2 holds for $\overline{\mathcal{R}} = \lambda\mathcal{R} + \delta_{\mathcal{C}}$. From the definition of $\mathcal{C}$ and $\mathcal{R}$, it follows that $\operatorname{dom}(\mathcal{R}) \cap \mathcal{C} = \mathcal{C}$, and subsequently, we have $\mathcal{D} := \mathcal{F} \cap \operatorname{dom}(\mathcal{R}) \cap \mathcal{C} = \mathcal{F} \cap \mathcal{C}$. One can easily see that $0 \in \mathcal{D}$. Since $\|\cdot\|_*$ is a norm on $\operatorname{dom}(\mathcal{R})$ and $\mathcal{C} \subset \operatorname{dom}(\mathcal{R})$, we know that $\mathcal{C}$ is a convex set, and also, $\mathcal{R}$ is a convex function. For $\mathrm{S} \in \mathcal{L}(\mathcal{H})$, we have the inequality $\|\mathrm{S}\| \leq \|\mathrm{S}\|_*$, which implies the coercivity of $\mathcal{R}$. Hence, $\overline{\mathcal{R}} = \lambda\mathcal{R} + \delta_{\mathcal{C}}$ is a coercive function. Let $\{\mathrm{S}_n\}_{n\in\mathbb{N}} \subset \mathcal{L}(\mathcal{H})$ be a sequence such that $\lim_{n\to\infty} \mathrm{S}_n = \mathrm{S} \in \mathcal{L}(\mathcal{H})$ in norm topology. Subsequently, we know that $\lim_{n\to\infty} |\mathrm{S}_n| = |\mathrm{S}|$ in norm topology. Hence, for any $k \in \mathbb{N}$, we have that $\lim_{n\to\infty} \langle b_k, |\mathrm{S}_n| b_k\rangle = \langle b_k, |\mathrm{S}| b_k\rangle$. Moreover, since $\{|\mathrm{S}_n|\}_{n\in\mathbb{N}}$ and $|\mathrm{S}|$ are nonnegative operators, we know that $\langle b_k, |\mathrm{S}_n| b_k\rangle \geq 0$ and $\langle b_k, |\mathrm{S}| b_k\rangle \geq 0$, for any $n, k \in \mathbb{N}$. Hence, from Fatou's lemma [45], it follows that

$$\begin{aligned}\|\mathrm{S}\|_* &= \textstyle\sum_{k=1}^{\infty} \langle b_k, |\mathrm{S}| b_k\rangle = \sum_{k=1}^{\infty} \liminf_n \langle b_k, |\mathrm{S}_n| b_k\rangle\\ &\leq \liminf_n \textstyle\sum_{k=1}^{\infty} \langle b_k, |\mathrm{S}_n| b_k\rangle = \liminf_n \|\mathrm{S}_n\|_*.\end{aligned} \tag{85}$$

Accordingly, we know that $\mathcal{R}(\mathrm{S}) \leq \liminf_n \mathcal{R}(\mathrm{S}_n)$, and $\mathcal{R}$ is lower semicontinuous. Also, if $\{\mathrm{S}_n\}_{n\in\mathbb{N}} \subset \mathcal{C}$, then $\|\mathrm{S}_n\|_* \leq \frac{1}{\lambda}\|\mathrm{Y}\|_{\mathrm{F}}$, for each n. Therefore, due to (85), we know that $\|\mathrm{S}\|_* \leq \frac{1}{\lambda}\|\mathrm{Y}\|_{\mathrm{F}}$, and, subsequently, we have $\mathrm{S} \in \mathcal{C}$. Hence, $\mathcal{C}$ and $\mathcal{D} = \mathcal{F} \cap \mathcal{C}$ are non-empty, closed and convex sets. Accordingly, Theorem 8 implies that (84) admits solution $\hat{\mathrm{K}}$ with the parametric form in (40), which is also a solution for (56) due to the equivalency of the corresponding programs. Let $\{b_n\}_{n\in\mathbb{N}}$ be an orthonormal basis such that $\mathcal{G} = \operatorname{span}\{b_j | j \in [\overline{n}_{\mathrm{g}}]\}$. Thus, for each $l \in [n_{\mathrm{g}}]$, there exist real numbers $e_{l1}, \ldots, e_{l\overline{n}_{\mathrm{g}}}$ such that $g_l = \sum_{j=1}^{\overline{n}_{\mathrm{g}}} e_{lj} b_j$. From (78), it follows that

$$\begin{aligned}\hat{\mathrm{K}}^*\hat{\mathrm{K}} &= \textstyle\sum_{j=1}^{n_{\mathrm{g}}} \sum_{l=1}^{n_{\mathrm{g}}} \left(\sum_{i=1}^{\overline{n}_{\mathrm{g}}} e_{li} b_i\right) \otimes \left(\sum_{k=1}^{\overline{n}_{\mathrm{g}}} e_{jk} b_k\right) [\mathrm{A}^{\mathsf{T}}\mathrm{ZA}]_{(l,j)}\\ &= \textstyle\sum_{i=1}^{\overline{n}_{\mathrm{g}}} \sum_{k=1}^{\overline{n}_{\mathrm{g}}} \left(\sum_{j=1}^{n_{\mathrm{g}}} \sum_{l=1}^{n_{\mathrm{g}}} e_{li} [\mathrm{A}^{\mathsf{T}}\mathrm{ZA}]_{(l,j)} e_{jk}\right) (b_i \otimes b_k)\\ &= \textstyle\sum_{i=1}^{\overline{n}_{\mathrm{g}}} \sum_{k=1}^{\overline{n}_{\mathrm{g}}} [\mathrm{E}^{\mathsf{T}}\mathrm{A}^{\mathsf{T}}\mathrm{ZAE}]_{(i,k)} (b_i \otimes b_k)\end{aligned}$$

where E is the matrix defined as $\mathrm{E} = [e_{lj}]_{l=1,j=1}^{n_{\mathrm{g}},\overline{n}_{\mathrm{g}}}$. One can easily see that $\mathrm{G} = \mathrm{EE}^{\mathsf{T}}$. Due to the equality above, we know that there exist r_{ik}, for $i, k \in [\overline{n}_{\mathrm{g}}]$, such that $|\hat{\mathrm{K}}^*\hat{\mathrm{K}}| = \sum_{i=1}^{\overline{n}_{\mathrm{g}}} \sum_{k=1}^{\overline{n}_{\mathrm{g}}} r_{ik} (b_i \otimes b_k)$. From (77), we have

$$\begin{aligned}|\hat{\mathrm{K}}^*\hat{\mathrm{K}}|^2 &= \textstyle\sum_{i=1}^{\overline{n}_{\mathrm{g}}} \sum_{j=1}^{\overline{n}_{\mathrm{g}}} r_{ij} (b_i \otimes b_j) \sum_{l=1}^{\overline{n}_{\mathrm{g}}} \sum_{k=1}^{\overline{n}_{\mathrm{g}}} r_{lk} (b_l \otimes b_k)\\ &= \textstyle\sum_{i=1}^{\overline{n}_{\mathrm{g}}} \sum_{k=1}^{\overline{n}_{\mathrm{g}}} \sum_{j=1}^{\overline{n}_{\mathrm{g}}} \sum_{l=1}^{\overline{n}_{\mathrm{g}}} r_{ij} r_{lk} \langle b_j, b_l\rangle (b_i \otimes b_k)\\ &= \textstyle\sum_{i=1}^{\overline{n}_{\mathrm{g}}} \sum_{k=1}^{\overline{n}_{\mathrm{g}}} [\mathrm{R}^2]_{(i,k)} (b_i \otimes b_k)\end{aligned}$$

where R is the matrix defined as $\mathrm{R} = [r_{ik}]_{i=1,k=1}^{\overline{n}_{\mathrm{g}},\overline{n}_{\mathrm{g}}}$. From the abovementioned calculations for $\hat{\mathrm{K}}^*\hat{\mathrm{K}}$ and $|\hat{\mathrm{K}}^*\hat{\mathrm{K}}|^2$, it follows that

$$\mathrm{R}^2 = \mathrm{E}^{\mathsf{T}}\mathrm{A}^{\mathsf{T}}\mathrm{ZAE} = \left(\mathrm{Z}^{\frac{1}{2}}\mathrm{AE}\right)^{\mathsf{T}} \left(\mathrm{Z}^{\frac{1}{2}}\mathrm{AE}\right). \tag{86}$$

Note that we have $\langle h_1, (h_2 \otimes h_3) h_1\rangle = \langle h_1, h_2\rangle\langle h_1, h_3\rangle$, for any $h_1, h_2, h_3, h_4 \in \mathcal{H}$. Therefore, we have

$$\begin{aligned}\|\hat{\mathrm{K}}\|_* &= \operatorname{tr}(|\hat{\mathrm{K}}^*\hat{\mathrm{K}}|) = \textstyle\sum_{j=1}^{\infty} \langle b_j, \sum_{i=1}^{\overline{n}_{\mathrm{g}}} \sum_{k=1}^{\overline{n}_{\mathrm{g}}} r_{ik} (b_i \otimes b_k) b_j\rangle\\ &= \textstyle\sum_{j=1}^{\infty} \sum_{i=1}^{\overline{n}_{\mathrm{g}}} \sum_{k=1}^{\overline{n}_{\mathrm{g}}} r_{ik} \langle b_j, b_i\rangle\langle b_j, b_k\rangle = \sum_{i=1}^{\overline{n}_{\mathrm{g}}} r_{ii} = \operatorname{tr}(\mathrm{R}).\end{aligned}$$

Accordingly, from (86), we know that $\|\hat{\mathrm{K}}\|_* = \|\mathrm{Z}^{\frac{1}{2}}\mathrm{AE}\|_*$. Note that, for any matrix M, we have $\|\mathrm{M}\|_* = \|\mathrm{M}^{\mathsf{T}}\|_* =$

$\mathrm{tr}((\mathrm{M}^\mathsf{T}\mathrm{M})^{\frac{1}{2}}) = \mathrm{tr}((\mathrm{M}\mathrm{M}^\mathsf{T})^{\frac{1}{2}})$. From this fact and $\mathrm{G}^{\frac{1}{2}}\mathrm{G}^{\frac{1}{2}} = \mathrm{G} = \mathrm{E}\mathrm{E}^\mathsf{T}$, it follows that

$$\begin{aligned}\|\hat{\mathrm{K}}\|_* &= \mathrm{tr}\left(\left(\mathrm{Z}^{\frac{1}{2}}\mathrm{A}\mathrm{E}\right)\left(\mathrm{E}^\mathsf{T}\mathrm{A}^\mathsf{T}\mathrm{Z}^{\frac{1}{2}}\right)\right)\\ &= \mathrm{tr}\left(\left(\mathrm{Z}^{\frac{1}{2}}\mathrm{A}\mathrm{G}^{\frac{1}{2}}\right)\left(\mathrm{G}^{\frac{1}{2}}\mathrm{A}^\mathsf{T}\mathrm{Z}^{\frac{1}{2}}\right)\right) = \|\mathrm{G}^{\frac{1}{2}}\mathrm{A}^\mathsf{T}\mathrm{Z}^{\frac{1}{2}}\|_*\\ &= \|\mathrm{Z}^{\frac{1}{2}}\mathrm{A}\mathrm{G}^{\frac{1}{2}}\|_*.\end{aligned}$$

Similar to the proof of Theorem 1, one can also show $\mathcal{E}(\hat{\mathrm{K}}) = \|\mathrm{Z}\mathrm{A}\mathrm{G} - \mathrm{Y}\|_\mathrm{F}^2$. Substituting these terms in (56), we obtain the optimization problem (52). ■

F. Proof of Theorem 13

Consider a trajectory of system (1) as $\{\mathrm{x}_n\}_{n\in\mathbb{N}}$, and let $g \in \mathcal{H}_\Bbbk$. From the reproducing property of kernel and Cauchy–Schwartz inequality, it follows that

$$|g(\mathrm{x}_n)| = |(\mathrm{K}^n g)(\mathrm{x}_0)| = |\langle \Bbbk(\mathrm{x}_0,\cdot), \mathrm{K}^n g\rangle| \le \|\Bbbk(\mathrm{x}_0,\cdot)\|\|\mathrm{K}^n g\|.$$

Since $\|\Bbbk(\mathrm{x}_0,\cdot)\| = \Bbbk(\mathrm{x}_0,\mathrm{x}_0)^{\frac{1}{2}}$ and $\|\mathrm{K}\| \le 1-\varepsilon$, we have $|g(\mathrm{x}_n)| \le (1-\varepsilon)^n \Bbbk(\mathrm{x}_0,\mathrm{x}_0)^{\frac{1}{2}}\|g\|$. Due to (60) and by replacing g with h_j, it follows that

$$L_j\,|x_{n,j} - x_{\mathrm{eq},j}|^{\alpha_j} \le |h_j(\mathrm{x}_n)| \le (1-\varepsilon)^n \Bbbk(\mathrm{x}_0,\mathrm{x}_0)^{\frac{1}{2}}\|h_j\|$$

where $x_{n,j}$ is the j^{th} coordinate of x_n, for $j \in [n_\mathrm{x}]$. Accordingly, for each j, we have

$$|x_{n,j} - x_{\mathrm{eq},j}| \le (1-\varepsilon)^{\frac{n}{\alpha}} \max_{j\in[n_\mathrm{x}]}\left[\frac{1}{L_j}\sup_{\mathrm{x}\in\mathcal{X}} \Bbbk(\mathrm{x},\mathrm{x})^{\frac{1}{2}}\|h_j\|\right]^{\frac{1}{\alpha_j}}$$

where $\alpha = \max\{\alpha_j | j \in [n_\mathrm{x}]\}$. Hence, we have $\lim_{n\to\infty}\mathrm{x}_n = \mathrm{x}_{\mathrm{eq}}$, where the convergence is uniform and with exponential rate. ■

G. Proof of Theorem 14

We know that $\|\Pi_\mathcal{Z}\mathrm{S}\Pi_\mathcal{G}\| \le \|\mathrm{S}\|$, for all $\mathrm{S} \in \mathcal{L}(\mathcal{H}_\Bbbk)$. Hence, $\|\mathrm{S}\| \le 1-\varepsilon$ implies that $\|\Pi_\mathcal{Z}\mathrm{S}\Pi_\mathcal{G}\| \le 1-\varepsilon$. Subsequently, we have $\delta_{\mathcal{C}_\varepsilon}(\Pi_\mathcal{Z}\mathrm{S}\Pi_\mathcal{G}) \le \delta_{\mathcal{C}_\varepsilon}(\mathrm{S})$, where $\mathcal{C}_\varepsilon$ is the nonempty, closed, and convex set defined as $\mathcal{C}_\varepsilon := \{\mathrm{S} \in \mathcal{L}(\mathcal{H}_\Bbbk) | \|\mathrm{S}\| \le 1-\varepsilon\}$. Thus, $\overline{\mathcal{R}}_\varepsilon := \delta_{\mathcal{C}_\varepsilon}$ satisfies Assumption 2, and claims are implied directly from Theorems 9 and 8. ■

H. Supporting Lemmas

Lemma 15: For matrices A and B, we have $\|\mathrm{A}\mathrm{B}\|_\mathrm{F} \le \|\mathrm{A}\|\|\mathrm{B}\|_\mathrm{F}$. Also, if B is invertible, then $\|\mathrm{A}\mathrm{B}\|_\mathrm{F} \le \|\mathrm{A}\|_\mathrm{F}\|\mathrm{B}\|$.

Lemma 16: The function $f : \mathcal{L}(\mathcal{H}) \to \mathbb{R}_+$, defined as $f(\mathrm{T}) = \|\mathrm{T}\|_\mathrm{F}^2$, is strictly convex.

Proof: Let $\{b_k\}_{k=1}^\infty$ be an orthonormal basis for $\mathcal{H}$, $t \in (0,1)$ and $\mathrm{T}_1, \mathrm{T}_2 \in \mathcal{L}(\mathcal{H})$ such that $\mathrm{T}_1 \ne \mathrm{T}_2$. Then, for each k, we have $\|\mathrm{T}_1 b_k\|^2 + \|\mathrm{T}_2 b_k\|^2 \ge 2\langle \mathrm{T}_1 b_k, \mathrm{T}_2 b_k\rangle$. The equality holds if and only if $\mathrm{T}_1 b_k = \mathrm{T}_2 b_k$. Since $\mathrm{T}_1 \ne \mathrm{T}_2$, there exists k such that this inequality is strict. Now, multiplying both sides with $t(1-t)$ and rearranging the terms, we have

$$\begin{aligned}t\|\mathrm{T}_1 b_k\|^2 + (1-t)\|\mathrm{T}_2 b_k\|^2 &\ge t^2\|\mathrm{T}_1 b_k\|^2 + (1-t)^2\|\mathrm{T}_2 b_k\|^2\\ &+ 2t(1-t)\langle \mathrm{T}_1 b_k, \mathrm{T}_2 b_k\rangle = \|(t\mathrm{T}_1 + (1-t)\mathrm{T}_2)b_k\|^2.\end{aligned}$$

By taking summation and due to the definition of Frobenius norm, we have $t\|\mathrm{T}_1\|_\mathrm{F}^2 + (1-t)\|\mathrm{T}_2\|_\mathrm{F}^2 > \|t\mathrm{T}_1 + (1-t)\mathrm{T}_2\|_\mathrm{F}^2$. Thus, it follows that $tf(\mathrm{T}_1) + (1-t)f(\mathrm{T}_2) > f(t\mathrm{T}_1 + (1-t)\mathrm{T}_2)$, which concludes the proof. ■

Lemma 17: For matrices $\mathrm{A} \in \mathbb{R}^{n\times m}$, $\mathrm{B} \in \mathbb{R}^{m\times k}$, and $\mathrm{C} \in \mathbb{R}^{k\times p}$, we have $\mathrm{rank}(\mathrm{AB}) + \mathrm{rank}(\mathrm{BC}) \le \mathrm{rank}(\mathrm{B}) + \mathrm{rank}(\mathrm{ABC})$.

Lemma 18: For matrices $\mathrm{A} \in \mathbb{R}^{n\times m}$, $\mathrm{B} \in \mathbb{R}^{m\times k}$, and $\mathrm{C} \in \mathbb{R}^{k\times p}$, we have $\mathrm{rank}(\mathrm{AB}) = \mathrm{rank}(\mathrm{A}^\mathsf{T}\mathrm{AB})$ and $\mathrm{rank}(\mathrm{BC}) = \mathrm{rank}(\mathrm{BCC}^\mathsf{T})$.

Proof: We know that $\mathrm{rank}(\mathrm{A}^\mathsf{T}\mathrm{AB}) \le \mathrm{rank}(\mathrm{AB})$. Therefore, from Lemma 18, it follows that $\mathrm{rank}(\mathrm{A}^\mathsf{T}\mathrm{A}) + \mathrm{rank}(\mathrm{AB}) \le \mathrm{rank}(\mathrm{A}) + \mathrm{rank}(\mathrm{A}^\mathsf{T}\mathrm{AB}) \le \mathrm{rank}(\mathrm{A}) + \mathrm{rank}(\mathrm{AB})$. Hence, from $\mathrm{rank}(\mathrm{A}^\mathsf{T}\mathrm{A}) = \mathrm{rank}(\mathrm{A})$, we have $\mathrm{rank}(\mathrm{AB}) = \mathrm{rank}(\mathrm{A}^\mathsf{T}\mathrm{AB})$. Similarly, we can show the other equality. ■

REFERENCES

[1] J. Schoukens and L. Ljung, "Nonlinear system identification: A user-oriented road map," *IEEE Control Syst. Mag.*, vol. 39, no. 6, pp. 28–99, Dec. 2019.

[2] J. F. Fisac, A. K. Akametalu, M. N. Zeilinger, S. Kaynama, J. Gillula, and C. J. Tomlin, "A general safety framework for learning-based control in uncertain robotic systems," *IEEE Trans. Autom. Control*, vol. 64, no. 7, pp. 2737–2752, Jul. 2019.

[3] H. Mania, M. I. Jordan, and B. Recht, "Active learning for nonlinear system identification with guarantees," *J. Mach. Learn. Res.*, vol. 23, pp. 1–32, 2022.

[4] E. Kaiser, J. N. Kutz, and S. L. Brunton, "Sparse identification of nonlinear dynamics for model predictive control in the low-data limit," *Proc. Roy. Soc. A*, vol. 474, no. 2219, 2018, Art. no. 20180335.

[5] J. Umlauft and S. Hirche, "Feedback linearization based on Gaussian processes with event-triggered online learning," *IEEE Trans. Autom. Control*, vol. 65, no. 10, pp. 4154–4169, Oct. 2019.

[6] S. M. Khansari-Zadeh and O. Khatib, "Learning potential functions from human demonstrations with encapsulated dynamic and compliant behaviors," *Auton. Robots*, vol. 41, no. 1, pp. 45–69, 2017.

[7] A. J. Ijspeert, J. Nakanishi, H. Hoffmann, P. Pastor, and S. Schaal, "Dynamical movement primitives: Learning attractor models for motor behaviors," *Neural Comput.*, vol. 25, no. 2, pp. 328–373, 2013.

[8] M. Khosravi and R. S. Smith, "Nonlinear system identification with prior knowledge on the region of attraction," *IEEE Control Syst. Lett.*, vol. 5, no. 3, pp. 1091–1096, Jul. 2021.

[9] A. A. Ahmadi and B. El Khadir, "Learning dynamical systems with side information (short version)," *Proc. Mach. Learn. Res.*, vol. 120, pp. 718–727, 2020.

[10] M. Khosravi and R. S. Smith, "Convex nonparametric formulation for identification of gradient flows," *IEEE Control Syst. Lett.*, vol. 5, no. 3, pp. 1097–1102, Jul. 2021.

[11] Y. S. Mauroy and I. Mezic, *Koopman Operator in Systems and Control*. Berlin, Germany: Springer, 2020.

[12] B. O. Koopman, "Hamiltonian systems and transformation in Hilbert space," *Proc. Nat. Acad. Sci. United States Amer.*, vol. 17, no. 5, pp. 315–318, 1931.

[13] R. K. Singh and J. S. Manhas, *Composition Operators on Function Spaces*. Amsterdam, The Netherlands: Elsevier, 1993.

[14] I. Mezić, "Spectral properties of dynamical systems, model reduction and decompositions," *Nonlinear Dyn.*, vol. 41, no. 1–3, pp. 309–325, 2005.

[15] I. Mezić and A. Banaszuk, "Comparison of systems with complex behavior," *Physica D, Nonlinear Phenomena*, vol. 197, no. 1/2, pp. 101–133, 2004.

[16] I. Mezić, "Spectrum of the Koopman operator, spectral expansions in functional spaces, and state-space geometry," *J. Nonlinear Sci.*, vol. 30, no. 5, pp. 1–55, 2019.

[17] S. L. Brunton, B. W. Brunton, J. L. Proctor, and J. N. Kutz, "Koopman invariant subspaces and finite linear representations of nonlinear dynamical systems for control," *PLoS One*, vol. 11, no. 2, 2016, Art. no. e0150171.

[18] E. Kaiser, J. N. Kutz, and S. L. Brunton, "Data-driven discovery of Koopman eigenfunctions for control," *Mach. Learn., Sci. Technol.*, vol. 2, no. 3, 2021, Art. no. 035023.

[19] G. Mamakoukas, M. L. Castano, X. Tan, and T. D. Murphey, "Derivative-based Koopman operators for real-time control of robotic systems," *IEEE Trans. Robot.*, vol. 37, no. 6, pp. 2173–2192, Dec. 2021.

[20] I. Abraham, G. De La Torre, and T. D. Murphey, "Model-based control using Koopman operators," *Proc. Robot., Sci. Syst.*, vol. 13, pp. 52, 2017.
[21] D. Bruder, B. Gillespie, C. D. Remy, and R. Vasudevan, "Modeling and control of soft robots using the Koopman operator and model predictive control," *Proc. Robotics, Sci. Syst.*, vol. 15, pp. 60, 2019.
[22] A. M. Boudali, P. J. Sinclair, R. Smith, and I. R. Manchester, "Human locomotion analysis: Identifying a dynamic mapping between upper and lower limb joints using the Koopman operator," in *Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*2017, pp. 1889–1892.
[23] B. W. Brunton, L. A. Johnson, J. G. Ojemann, and J. N. Kutz, "Extracting spatial–temporal coherent patterns in large-scale neural recordings using dynamic mode decomposition," *J. Neurosci. Methods*, vol. 258, pp. 1–15, 2016.
[24] I. Mezić, "Analysis of fluid flows via spectral properties of the Koopman operator," *Annu. Rev. Fluid Mech.*, vol. 45, pp. 357–378, 2013.
[25] J. Hogg, M. Fonoberova, and I. Mezić, "Exponentially decaying modes and long-term prediction of sea ice concentration using Koopman mode decomposition," *Sci. Rep.*, vol. 10, no. 1, pp. 1–15, 2020.
[26] C. Folkestad, D. Pastor, I. Mezic, R. Mohr, M. Fonoberova, and J. Burdick, "Extended dynamic mode decomposition with learned Koopman eigenfunctions for prediction and control," in *Proc. IEEE Amer. Control Conf.*, 2020, pp. 3906–3913.
[27] I. Abraham and T. D. Murphey, "Active learning of dynamics for data-driven control using Koopman operators," *IEEE Trans. Robot.*, vol. 35, no. 5, pp. 1071–1083, Oct. 2019.
[28] G. Mamakoukas, I. Abraham, and T. D. Murphey, "Learning stable models for prediction and control," 2020, *arXiv:2005.04291*.
[29] K. Hara, M. Inoue, and N. Sebe, "Learning Koopman operator under dissipativity constraints," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 1169–1174, 2020.
[30] N. Takeishi, Y. Kawahara, and T. Yairi, "Learning Koopman invariant subspaces for dynamic mode decomposition," in *Proc. Neural Inf. Process. Syst.*, 2017, pp. 1130–1140.
[31] M. Haseli and J. Cortés, "Learning Koopman eigenfunctions and invariant subspaces from data: Symmetric subspace decomposition," *IEEE Trans. Autom. Control*, vol. 67, no. 7, pp. 3442–3457, Jul. 2022.
[32] P. J. Schmid, "Dynamic mode decomposition of numerical and experimental data," *J. Fluid Mechanics*, vol. 656, pp. 5–28, 2010.
[33] H. Arbabi and I. Mezic, "Ergodic theory, dynamic mode decomposition, and computation of spectral properties of the Koopman operator," *SIAM J. Appl. Dynamical Syst.*, vol. 16, no. 4, pp. 2096–2126, 2017.
[34] 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.
[35] S. Klus, F. Nüske, S. Peitz, J.-H. Niemann, C. Clementi, and C. Schütte, "Data-driven approximation of the Koopman generator: Model reduction, system identification, and control," *Physica D, Nonlinear Phenomena*, vol. 406, 2020, Art. no. 132416.
[36] B. Schölkopf, R. Herbrich, and A. J. Smola, "A generalized representer theorem," in *Proc. Int. Conf. Comput. Learn. Theory*., 2001, pp. 416–426.
[37] F. Dinuzzo and B. Schölkopf, "The representer theorem for Hilbert spaces: A necessary and sufficient condition," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, pp. 189–196, 2012.
[38] M. Unser, "A representer theorem for deep neural networks," *J. Mach. Learn. Res.*, vol. 20, no. 110, pp. 1–30, 2019.
[39] M. Khosravi, "Learning finite-dimensional representations for Koopman operators," in *Proc. Learn. Dyn. Control.*, 2021, pp. 1281–1281.
[40] N. Aronszajn, "Theory of reproducing kernels," *Trans. Amer. Math. Soc.*, vol. 68, no. 3, pp. 337–404, 1950.
[41] A. Berlinet and C. Thomas-Agnan, *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Berlin, Germany: Springer, 2011.
[42] H. Brezis, *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Berlin, Germany: Springer, 2010.
[43] J. Peypouquet, *Convex Optimization in Normed Spaces: Theory, Methods and Examples*. Berlin, Germany: Springer, 2015.
[44] M. Khosravi, "Representer theorem for learning Koopman operators," 2022, *arXiv:2208.01681*.
[45] C. D. Aliprantis and K. C. Border, *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Berlin, Germany: Springer, 2006.
[46] S. Peitz and S. Klus, "Koopman operator-based model reduction for switched-system control of PDEs," *Automatica*, vol. 106, pp. 184–191, 2019.
[47] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank," *Psychometrika*, vol. 1, no. 3, pp. 211–218, 1936.
[48] S. Ji and J. Ye, "An accelerated gradient method for trace norm minimization," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 457–464.
[49] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Rev.*, vol. 52, no. 3, pp. 471–501, 2010.
[50] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA, USA: MIT Press, 2006.
[51] N. F. Britton, *Essential Mathematical Biology*, vol. 453. Berlin, Germany: Springer, 2003.

Mohammad Khosravi (Member, IEEE) received the B.Sc. degree in electrical engineering and the B.Sc. degree in mathematical sciences from the Sharif University of Technology, Tehran, Iran, in 2011, and the postgraduate diploma in mathematics from The Abdus Salam International Centre for Theoretical Physics, Trieste, Italy, in 2012, and the M.A.Sc. degree in electrical and computer engineering from Concordia University, Montreal, QC, Canada, in 2016, and the Ph.D. degree in information technology and electrical engineering from the Swiss Federal Institute of Technology (ETH), Zürich, Switzerland, in 2022.

He is currently an Assistant Professor with the Delft Center for Systems and Control (DCSC), Delft University of Technology, Delft, The Netherlands. He was a Research Assistant with the Mathematical Biology Group, Institute for Research in Fundamental Sciences, Tajrish, Iran, from 2012 to 2014. His research interests involve data-driven and learning-based methods in modeling, model reduction, optimization, and control of dynamical systems and their applications in thermodynamics, buildings, energy, industry, and power systems.

Dr. Khosravi was the recipient of several awards, including the Gold Medal of the National Mathematics Olympiad, the Outstanding Student Paper Award in CDC 2020, the Silver Medal of ETH Zürich, and the Outstanding Reviewer Award for IEEE JOURNAL OF CONTROL SYSTEMS LETTERS.