# Sparse Learning of Kernel Transfer Operators

Boya Hou Subhonmesh Bose Umesh Vaidya

***Abstract*—Transfer operators such as the Koopman and the Perron-Frobenius operators provide valuable insights into the properties of nonlinear dynamical systems. Recent work has shown that non-parametric approximations of these operators can be constructed over reproducing kernel Hilbert space (RKHS) with data. These kernel transfer operators can then be written as functions of covariance and cross-covariance operators associated with the data generated by the dynamical system. In this paper, we study *sparse* kernel learning methods for kernel transfer operators. Specifically, we study sample complexity guarantees for coherency-based sparsification and demonstrate its efficacy over an example dynamical system.**

## I. Introduction

Transfer operators such as the Koopman and the Perron-Frobenius operators have emerged as powerful tools to analyze global behavior of nonlinear dynamical systems. These operators essentially *lift* the nonlinear dynamical system description over a finite-dimensional state space to a linear infinite-dimensional description that captures the action of the system dynamics on suitable spaces of functions. This approach allows us to carry over mature intuitions from linear systems theory to the study of nonlinear systems. The spectra of these operators are rich in information; they can be used to decompose modes of a dynamical system, propagate uncertainties and analyze global stability of the dynamics among other application uses.

Computational techniques have been widely studied to obtain finite-dimensional approximations of these infinite-dimensional operators from data. One can study the approximate spectra by studying the actions of these operators on parameterized function spaces, e.g., using the so-called extended dynamic mode decomposition in [1]. Even neural networks have been utilized to parameterize these function spaces, e.g., see [2]. Another line of research has sought to study the interactions of these operators with reproducing kernel Hilbert spaces (RKHS) of functions. See [3] for the use of kernel methods to approximate Koopman operators from data. Non-parametric methods in data science have a long history (e.g., see [4]) and offer powerful tools that, through transfer operators, can now be utilized to analyze nonlinear dynamical systems.

The recent work in [5] reveals deep connections between the transfer operators and the widely studied covariance and cross-covariance operators on RKHS. See [6] for a reference. As a result, sample complexities for data-driven approximations of these transfer operators follow from well-known convergence properties of covariance/cross-covariance operators in [7], [8].

Studying the actions of operators in RKHS can prove computationally burdensome with large volumes of data. This downside of learning over RKHS has given rise to a rich literature on *sparsification* that seeks to "throw away" those data points that do not add enough extra information to those obtained from the other data points, as in [9]–[11]. Sparsification in kernel methods is crucial for scalability. In this paper, we study the impact of sparsification in learning of kernel transfer operators. Specifically, we consider sparse kernel learning that utilizes the notion of *coherency* to control the growth of data points. We provide sample complexity guarantees for such sparsification for learning of kernel transfer operators and numerically illustrate its impact on the spectra of an example dynamical system.

## II. RKHS Preliminaries

We begin by formally defining a reproducing kernel Hilbert space (RKHS). See [6] for an introduction. Let $\mathbb{X}$ be a compact subset of an Euclidean space and $\kappa : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ be a continuous, symmetric, positive semi-definite kernel. Define $\mathcal{H}$ as the RKHS associated with the kernel $\kappa$–the completion of the span of $\{\phi(x) := \kappa(x, \cdot) : x \in \mathbb{X}\}$, equipped with the inner product $\langle \cdot, \cdot \rangle$, satisfying $\langle \phi(x), \phi(y) \rangle = \kappa(x, y)$. Here, $\phi$ is called the feature map for kernel $\kappa$. The inner product satisfies the reproducing property, given by

$$\langle \phi(x), f \rangle = f(x), \quad \forall x \in \mathbb{X}, \quad f \in \mathcal{H}. \tag{1}$$

Probability measures over $\mathbb{X}$ can be embedded within an RKHS. For a probability space $(\mathbb{X}, \Sigma, \mathbb{P})$, with Borel $\sigma$-algebra $\Sigma$, the *kernel mean embedding* of $\mathbb{P}$ in $\mathcal{H}$ is

$$\mu_{\mathbb{P}} := \mathbb{E}\left[\kappa(X, \cdot)\right] \quad \text{for } X \sim \mathbb{P}. \tag{2}$$

Assume throughout that $\kappa$ is measurable. If the kernel is finite, i.e., $\mathbb{E}[\kappa(X, X)] < \infty$, then $\mu_{\mathbb{P}} \in \mathcal{H}$, according to [6, Lemma 3.1]. Thus, the probability measure $\mathbb{P}$ is identified as an element in the RKHS.

With a slight abuse of notation, if $\mathbb{P}(X, Y)$ denotes a joint distribution over $\mathbb{X} \times \mathbb{X}$, then $\mathbb{P}$ can be embedded in the tensor product space $\mathcal{H} \otimes \mathcal{H}$, per [12], as

$$C_{XY} := \mathbb{E}_{XY}[\phi(X) \otimes \phi(Y)] = \mu_{\mathbb{P}_{XY}}. \tag{3}$$

$\mathcal{H} \otimes \mathcal{H}$ is equipped with the kernel $\kappa_{\otimes}$, defined by

$$\kappa_{\otimes}\Big((x_1, y_1), (x_2, y_2)\Big) = \kappa(x_1, x_2)\ \kappa(y_1, y_2) \tag{4}$$

for $x_1, x_2, y_1, y_2$ in $\mathbb{X}$. Its joint feature map is

$$\varphi\left(x_i, y_i\right) := \phi\left(x_i\right) \otimes \phi\left(y_i\right) = \kappa\left(x_i, \cdot\right) \kappa\left(y_i, \cdot\right). \tag{5}$$

B. Hou and S. Bose are with the Department of Electrical and Computer Engineering at the University of Illinois at Urbana-Champaign, Urbana, IL 61801. U. Vaidya is with the Department of Mechanical Engineering in Clemson University, Clemson, SC 29634. This work was partially supported by the NSF-EPCN-2031570 grant.

The above relation also provides an interpretation for the tensor product in (3). The feature map $\varphi$ satisfies

$$\langle \varphi(x_i, y_i), \varphi(x_j, y_j) \rangle = \kappa(x_i, x_j)\, \kappa(y_i, y_j). \tag{6}$$

In (3), we identify $C_{XY}$ as an element in the tensor product space. It can also be viewed as a Hilbert-Schmidt (HS) linear operator $C_{XY} : \mathcal{H} \to \mathcal{H}$ that satisfies

$$\mathbb{E}_{XY}[f(X)g(Y)] = \langle C_{XY} g, f \rangle, \quad \forall f, g \in \mathcal{H}. \tag{7}$$

$C_{XY}$ is commonly known as the (uncentered) *cross-covariance* operator. For details on HS operators, see [6, Chapter 2.3]. Similarly, one can also define the (uncentered) covariance operator as

$$C_{XX} := \mathbb{E}_X[\phi(X) \otimes \phi(X)], \tag{8}$$

that can be viewed as the embedding of the marginal distribution $\mathbb{P}(X)$ in $\mathcal{H} \otimes \mathcal{H}$. In our setting, $\mathbb{E}[\kappa(X, X)] < \infty$ ensures that $C_{XX}$ and $C_{XY}$ are bounded. The former is also self-adjoint (see [13, Theorem 1]).

For a given kernel $\kappa$, we call $\mathcal{H}$ a *separable* Hilbert space if it admits a countable basis. With $\kappa$ being a continuous kernel over compact $\mathbb{X}$, $\mathcal{H}$ can indeed be shown to be separable, according to [14, Lemma 4.33].

## III. Defining Kernel Transfer Operators Using the Conditional Mean Embedding Operator

Consider a discrete-time stochastic dynamical system on state-space $\mathbb{X}$, described by the transition kernel $p$ as

$$\mathbb{P}\{x_{t+1} \in \mathbb{A} | x_t = x\} = \int_{\mathbb{A}} p(y|x) \mathrm{d}y, \tag{9}$$

for $\mathbb{A} \subseteq \mathbb{X}$, where $x_t$ denotes the state at time $t$. If $f$ is a probability density over $\mathbb{X}$, then the Perron–Frobenius operator $\mathcal{P}$ propagates $f$ through the system dynamics as

$$(\mathcal{P}f)(y) = \int p(y|x) f(x) \mathrm{d}x. \tag{10}$$

If $f$ is an observable (scalar-valued map) over $\mathbb{X}$, then the Koopman operator $\mathcal{K}$ acts on $f$ as

$$(\mathcal{K}f)(x) = \int p(y|x) f(y) \mathrm{d}y. \tag{11}$$

These transfer operators are infinite-dimensional but linear. The stochastic nonlinear propagation of a finite-dimensional state, described by the transition kernel $p$ can be studied via the linear propagation of functions by these infinite-dimensional operators. The spectra of the Perron–Frobenius (P-F) and the Koopman operator can be utilized to characterize basins of attraction, perform model reduction, propagate uncertainties and analyze global stability of the dynamics among other application uses. We study these operators where they interact with an RKHS. Specifically, we relate these operators to the conditional mean embedding operator, defined in [15].

Consider the joint distribution $\mathbb{P}(X, Y)$ over $\mathbb{X} \times \mathbb{X}$, where $X$ is sampled according to a reference distribution and $Y$ is sampled from a one-step propagation of $X$ through the system dynamics in (9) according to the transition kernel $p$. Then, the mean embedding of the conditional distribution $P(Y|x)$ into $\mathcal{H}$ is given by

$$\mu_{Y|x} := \mathbb{E}_{Y|x}[\phi(Y)|X = x] \tag{12}$$

for $x \in \mathbb{X}$. The conditional mean embedding operator $\mathcal{U}_{Y|X} : \mathcal{H} \to \mathcal{H}$, according to [15], is a linear operator that satisfies

$$\mu_{Y|x} = \mathcal{U}_{Y|X} \phi(x). \tag{13}$$

Then, we have

$$\begin{aligned} \mu_Y &= \mathbb{E}_Y[\phi(Y)] \\ &\overset{(a)}{=} \mathbb{E}_X[\mathbb{E}_{Y|X}[\phi(Y)|X]] \\ &\overset{(b)}{=} \mathbb{E}_X[\mu_{Y|X}] \\ &\overset{(c)}{=} \mathbb{E}_X[\mathcal{U}_{Y|X}\phi(X)] \\ &\overset{(d)}{=} \mathcal{U}_{Y|X}\mathbb{E}_X[\phi(X)] \\ &= \mathcal{U}_{Y|X}\mu_X. \end{aligned} \tag{14}$$

In the above derivation, (a) follows from the law of total expectation. Lines (b) and (c) are consequences of (12) and (13), respectively. Line (d) follows from the linearity $\mathcal{U}_{Y|X}$. The above relation motivates the definition of the embedded P-F operator, per [5], as $\mathcal{P} := \mathcal{U}_{Y|X}$ that propagates the embedded distribution of states through the system dynamics. Under the assumption that $\mathbb{E}[f(Y)|X] \in \mathcal{H}$ for all $f \in \mathcal{H}$, it follows from [15, Theorem 4] that

$$\mathcal{P} = \mathcal{U}_{Y|X} = C_{YX} C_{XX}^{-1}. \tag{15}$$

To identify the Koopman operator in terms of the conditional mean embedding operator, note that

$$\begin{aligned} \langle \mathcal{K}f, \phi(x) \rangle &\overset{(a)}{=} (\mathcal{K}f)(x) \\ &\overset{(b)}{=} \mathbb{E}[f(Y)|X = x] \\ &\overset{(c)}{=} \langle f, \mu_{Y|x} \rangle \\ &\overset{(d)}{=} \langle f, \mathcal{U}_{Y|X}\phi(x) \rangle \end{aligned} \tag{16}$$

for all $f \in \mathcal{H}$. Lines (a)-(d) follow respectively, from the reproducing property of $\kappa$, (11), (12), and (13). Thus, we identify the kernel Koopman operator $\mathcal{K}$ as the adjoint of $\mathcal{U}_{Y|X} = \mathcal{P}$, given by

$$\mathcal{K} := C_{XX}^{-1} C_{XY}. \tag{17}$$

The above definitions of $\mathcal{P}$ and $\mathcal{K}$ rely on $C_{XX}$ being an invertible map. For technical reasons, Klus et al. in [5] consider their regularized versions, defined as

$$\mathcal{P}_\varepsilon := C_{YX} (C_{XX} + \varepsilon I)^{-1}, \tag{18}$$

$$\mathcal{K}_\varepsilon := (C_{XX} + \varepsilon I)^{-1} C_{XY} \tag{19}$$

for $\varepsilon > 0$, where $I$ is the identity operator. In the sequel, we study data-driven sparse kernel approximation of the kernel Koopman operator $\mathcal{K}$. Guarantees for approximations of $\mathcal{P}$ can be similarly derived.

### A. Approximating the Kernel Koopman Operator with Data

Given $m$ data points $\mathcal{M} := \{(x_1, y_1), \ldots, (x_m, y_m)\}$ sampled i.i.d. from $\mathbb{P}(X, Y)$, one can compute empirical estimates of $C_{XX}$ and $C_{XY}$, respectively as

$$\begin{aligned} \widetilde{C}_{XX} &= \frac{1}{m}\sum_{i=1}^{m} \phi(x_i) \otimes \phi(x_i) = \frac{1}{m}\sum_{i=1}^{m} \varphi(x_i, x_i), \\ \widetilde{C}_{XY} &= \frac{1}{m}\sum_{i=1}^{m} \phi(x_i) \otimes \phi(y_i) = \frac{1}{m}\sum_{i=1}^{m} \varphi(x_i, y_i). \end{aligned} \tag{20}$$

Approximation of transfer operators in RKHS was analyzed in [5]. Their work showed that, with regularization parameter $\varepsilon$, the empirical estimator constructed using $m$ samples converges in operator norm at a rate $\mathcal{O}_p\left(m^{-1/2}\varepsilon^{-1}\right)$.[1] In this work, we propose *sparse* learning of kernel Koopman operator that only utilizes a subset of the data, in effect making the empirical estimation memory-efficient. We show that the generated sparse estimator converges to a neighborhood of the true kernel Koopman operator. Moreover, the extent of approximation can be controlled via a tunable parameter. Specifically, this parameter allows one to trade-off between sparsity (data-efficiency) and accuracy of approximation.

## IV. Sparse Learning of Kernel Koopman Operator

When the total number of data points $m$ grows large, computation of the empirical kernel Koopman operators becomes increasingly difficult. The challenge arises from the fact that each data-point $(x, y)$ adds a new kernel function centered around the new data point in the computation of the empirical operators. Such difficulties in kernel learning are well-documented, e.g., see [11], [16]–[19]. To circumvent this difficulty, we propose to prune $\mathcal{M}$ to construct a sparse dictionary $\mathcal{D}$ and then represent the empirical kernel transfer operators using the data in $\mathcal{D}$.

We employ and analyze dictionary sparsification based on the notion of *coherency* (see [10]). The key idea is to retain only those points $(x, y)$ in the dictionary that are not "too similar", where similarity is measured via the kernel $\kappa_\otimes$. For a given dataset $\mathcal{M}$, we construct $\mathcal{D}$ by identifying a subset of $\mathcal{M}$ that satisfies

$$\frac{\left|\kappa_\otimes\Big((x_i, y_i), (x_j, y_j)\Big)\right|}{\sqrt{\kappa_\otimes\Big((x_i, y_i), (x_i, y_i)\Big)\kappa_\otimes\Big((x_j, y_j), (x_j, y_j)\Big)}} \leq \gamma, \tag{21}$$

for each $i, j$ such that $(x_i, y_i), (x_j, y_j)$ are in $\mathcal{D}$. One can construct such a $\mathcal{D}$ as follows. Compute the Gram matrix using all elements in $\mathcal{M}$ with $\kappa_\otimes$ as the kernel. If $(x_s, y_s)$ and $(x_t, y_t)$ are such that they violate (21) with indices $s < t$, retain $(x_s, y_s)$ in $\mathcal{D}$, but discard $(x_t, y_t)$ from $\mathcal{D}$ and remove the row/column associated with $(x_t, y_t)$ from the Gram matrix. Repeat this operation until all elements in the Gram matrix satisfies (21). Let $\mathcal{I}$ be the indices among $1, \ldots, m$ for which $(x_i, y_i)$ are in $\mathcal{D}$. Then, the sparse estimator of $C_{XY}$ ($C_{XX}$) is:

$$\widehat{C}_{XY} = \sum_{i\in\mathcal{I}} \alpha_i \varphi(x_i, y_i), \quad \widehat{C}_{XX} = \sum_{i\in\mathcal{I}} \beta_i \varphi(x_i, x_i). \tag{22}$$

where $\alpha$ (and similarly, $\beta$) is defined as

$$\alpha := \operatorname*{argmin}_{\overline{\alpha}} \left\| \frac{1}{m}\sum_{i=1}^{m} \varphi(y_i, x_i) - \sum_{i\in\mathcal{I}} \overline{\alpha}_i \varphi(y_i, x_i) \right\|_{\mathcal{H}}^2, \tag{23}$$

The vector $\alpha$ admits the explicit representation $\alpha = G^{-1}g$. Here, $G \in \mathbb{R}^{|\mathcal{D}|\times|\mathcal{D}|}$ is the Gram matrix associated with elements in $\mathcal{D}$, given by

$$G_{i,j} = \kappa_\otimes\Big((x_i, y_i), (x_j, y_j)\Big) \tag{24}$$

for each $i$ and $j$ in $\mathcal{I}$ and $g \in \mathbb{R}^{|\mathcal{I}|}$ is defined as

$$[g]_j = \frac{1}{m}\sum_{i=1}^{m} \kappa_\otimes\Big((x_i, y_i), (x_j, y_j)\Big) \tag{25}$$

for each $j$ in $\mathcal{I}$. Using the sparse covariance operator $\widehat{C}_{XX}$ and cross-covariance operator $\widehat{C}_{XY}$, we define the *sparse kernel Koopman estimator* as

$$\widehat{\mathcal{K}}_\varepsilon := \left(\widehat{C}_{XX} + \varepsilon I\right)^{-1} \widehat{C}_{XY}. \tag{26}$$

We now bound the approximation error of this sparse kernel Koopman estimate in our main result.

**Theorem 1.** *Suppose $\kappa$ is a continuous kernel defined over a compact set $\mathbb{X}$. Then, $\widehat{\mathcal{K}}_\varepsilon$ in (26) and $\mathcal{K}_\varepsilon$ in (19) satisfies*

$$\left\|\widehat{\mathcal{K}}_\varepsilon - \mathcal{K}_\varepsilon\right\| \leq \psi(m, \gamma; \delta)\mathcal{O}(\varepsilon^{-2}), \tag{27}$$

*with probability at least $1 - \delta$ for $\delta \in (0, 1)$, where*

$$\begin{aligned} \psi(m, \gamma; \delta) := \frac{1}{\sqrt{m}}\left(1 + \sqrt{2\log(1/\delta)}\right) \\ + \left(1 - \frac{|\mathcal{D}|}{m}\right)\sqrt{1 - \gamma^2}. \end{aligned} \tag{28}$$

The error between the sparse estimator and the regularized operator is measured in *operator norm*, where $\|\mathcal{A}\| := \sup_{\|f\|=1} \|\mathcal{A}f\|$ for an operator $\mathcal{A}$. The above result indicates that the performance depends not only on the number of training samples $m$, but also on the coherence parameter $\gamma$. Compared with [5, Theorem 3.14], the second summand in $\psi$ arises due to sparsification. We use a technique inspired by [20] to derive the approximation error from sparsification. Upon decreasing $\gamma$, we obtain a less coherent dictionary with smaller $\mathcal{D}$. Our result shows that data sparsity comes at the expense of approximation accuracy for the kernel Koopman operator. As $m$ grows without bound, one can show that $|\mathcal{D}|$ saturates at some point (see [10, Proposition 2]). That is, the dictionary remains finite, even though the number of samples grows infinitely large. Thus, for large $m$, we have $\psi \sim \sqrt{1 - \gamma^2}$ that captures the price we pay for sparsification. We cannot avoid this cost with more data.

[1] We believe the dependence on $\varepsilon$ should be $\varepsilon^{-2}$ rather than $\varepsilon^{-1}$.

### A. Proof Sketch of Theorem 1

We utilize an argument similar to that in [5] to obtain

$$\begin{aligned}\left\|\widehat{\mathcal{K}}_\varepsilon - \mathcal{K}_\varepsilon\right\| \leq & \frac{1}{\varepsilon}\left\|\widehat{C}_{XY} - C_{XY}\right\|_{\text{HS}} \\ & + \frac{1}{\varepsilon^2}\left\|C_{XX} - \widehat{C}_{XX}\right\|_{\text{HS}} \|C_{XY}\|,\end{aligned} \tag{29}$$

where $\|\cdot\|_{\text{HS}}$ denotes the Hilbert-Schmidt norm.

Then, we bound the errors in cross-covariance estimation as a sum of two terms in

$$\left\|\widehat{C}_{XY} - C_{XY}\right\|_{\text{HS}} \leq \underbrace{\left\|\widetilde{C}_{XY} - C_{XY}\right\|_{\text{HS}}}_{\text{Sampling error}} + \underbrace{\left\|\widehat{C}_{XY} - \widetilde{C}_{XY}\right\|_{\text{HS}}}_{\text{Sparsification error}}. \tag{30}$$

We tackle the two terms separately. Precisely, we use [21] to bound the sampling error with probability $1-\delta$ as

$$\left\|\widetilde{C}_{XY} - C_{XY}\right\|_{\text{HS}} \leq \sqrt{B/m}\left(1 + \sqrt{2\log(1/\delta)}\right), \tag{31}$$

where $B := \sup_{x\in\mathbb{X}} \kappa^2(x,x)$.

The result then follows from bounding the sparsification error as

$$\left\|\widehat{C}_{XY} - \widetilde{C}_{XY}\right\|_{\text{HS}} \leq (1 - |\mathcal{D}|/m)\sqrt{B(1-\gamma^2)}, \tag{32}$$

using an argument similar to that in [20].

## V. Proof of Theorem 1

Boya says: This section is not included in the conference paper Triangle inequality and elementary algebra gives

$$\begin{aligned}&\left\|\widehat{\mathcal{K}}_\varepsilon - \mathcal{K}_\varepsilon\right\| \\ &\leq \left\|\left(\widehat{C}_{XX} + \varepsilon I\right)^{-1}\widehat{C}_{XY} - \left(\widehat{C}_{XX} + \varepsilon I\right)^{-1} C_{XY}\right\| \\ &\quad + \left\|\left(\widehat{C}_{XX} + \varepsilon I\right)^{-1} C_{XY} - (C_{XX} + \varepsilon I)^{-1} C_{XY}\right\| \\ &= \left\|\left(\widehat{C}_{XX} + \varepsilon I\right)^{-1}\left(\widehat{C}_{XY} - C_{XY}\right)\right\| \\ &\quad + \left\|\left[\left(\widehat{C}_{XX} + \varepsilon I\right)^{-1} - (C_{XX} + \varepsilon I)^{-1}\right] C_{XY}\right\|.\end{aligned} \tag{33}$$

Call the two norms in the last line as $Z_1$ and $Z_2$, respectively. We now bound $Z_1$ and $Z_2$ separately. Denote by $\|\cdot\|_{\text{HS}}$, the Hilbert-Schmidt norm of a bounded linear operator. We upper bound $Z_1$ as

$$\begin{aligned}Z_1 &\overset{(a)}{\leq} \left\|\left(\widehat{C}_{XX} + \varepsilon I\right)^{-1}\right\| \left\|\widehat{C}_{XY} - C_{XY}\right\| \\ &\overset{(b)}{\leq} \left\|\left(\widehat{C}_{XX} + \varepsilon I\right)^{-1}\right\| \left\|\widehat{C}_{XY} - C_{XY}\right\|_{\text{HS}} \\ &\overset{(c)}{\leq} \frac{1}{\varepsilon}\left\|\widehat{C}_{XY} - C_{XY}\right\|_{\text{HS}}.\end{aligned} \tag{34}$$

Here, (a) follows from the submultiplicative nature of the operator norm. Inequality (b) follows from the fact that the operator norm is dominated by the Hilbert-Schmidt norm. To get (c), note the the covariance operator $C_{XX}$ and its empirical estimate $\widehat{C}_{XX}$ are self-adjoint and positive semi-definite [2], implying that

$$\left\|\left(\widehat{C}_{XX} + \varepsilon I\right)^{-1}\right\| \leq \frac{1}{\varepsilon}. \tag{35}$$

Proceeding similarly, we bound $Z_2$ as

$$\begin{aligned}Z_2 &\leq \left\|\left(\widehat{C}_{XX} + \varepsilon\mathcal{I}\right)^{-1} - (C_{XX} + \varepsilon\mathcal{I})^{-1}\right\| \|C_{XY}\| \\ &= \left\|\left(\widehat{C}_{XX} + \varepsilon I\right)^{-1}\left(C_{XX} - \widehat{C}_{XX}\right)(C_{XX} + \varepsilon\mathcal{I})^{-1}\right\| \\ &\quad \times \|C_{XY}\| \\ &\leq \frac{1}{\varepsilon^2}\left\|C_{XX} - \widehat{C}_{XX}\right\| \|C_{XY}\| \\ &\leq \frac{1}{\varepsilon^2}\left\|C_{XX} - \widehat{C}_{XX}\right\|_{\text{HS}} \|C_{XY}\|.\end{aligned} \tag{36}$$

Here, the second last line follows from using the relation (35) and its counterpart with $\widehat{C}_{XX}$ replaced by $C_{XX}$. Utilizing (34) and (36) in (33), we get

$$\begin{aligned}\left\|\widehat{\mathcal{K}}_\varepsilon - \mathcal{K}_\varepsilon\right\| \leq & \frac{1}{\varepsilon}\left\|\widehat{C}_{XY} - C_{XY}\right\|_{\text{HS}} \\ & + \frac{1}{\varepsilon^2}\left\|C_{XX} - \widehat{C}_{XX}\right\|_{\text{HS}} \|C_{XY}\|.\end{aligned} \tag{37}$$

Recall that $\widehat{C}$ stands for the sparse empirical estimates, while $\widetilde{C}$ encodes the non-sparse empirical estimates of the covariance and cross-covariance operators. Using triangle inequality, we get

$$\left\|\widehat{C}_{XY} - C_{XY}\right\|_{\text{HS}} \leq \left\|\widehat{C}_{XY} - \widetilde{C}_{XY}\right\|_{\text{HS}} + \left\|\widetilde{C}_{XY} - C_{XY}\right\|_{\text{HS}}. \tag{38}$$

The same holds for $C_{XX}, \widehat{C}_{XX}$ and $\widetilde{C}_{XX}$. The rest of the proof bounds the two terms on the right-hand-side of (38).

First, note that the HS norm of (cross) covariance operator from $\mathcal{H}$ to $\mathcal{H}$ is the $\mathcal{H}_\otimes$-norm of the operator viewed as an element in tensor product Hilbert space $\mathcal{H}\otimes\mathcal{H}$ [7, Lemma4]:

$$\|C_{YX}\|_{\text{HS}}^2 = \|\mathbb{E}_{YX}\left[\kappa(\cdot,Y)\kappa(\cdot,X)\right]\|_{\mathcal{H}_\otimes}^2 \tag{39}$$

Moreover, recall that $C_{XY}$ and $C_{XX}$ are the embeddings of $\mathbb{P}(X,Y)$ and its marginal $\mathbb{P}(X)$ in $\mathcal{H}\otimes\mathcal{H}$. Using [6, Theorem 3.4], we bound the first term on the right-hand-side of (38) as

$$\left\|\widetilde{C}_{XY} - C_{XY}\right\|_{\text{HS}} \leq \sqrt{B/m}\left(1 + \sqrt{2\log(1/\delta)}\right) \tag{40}$$

with probability at least $1-\delta$, where $B := \sup_{x\in\mathbb{X}} \kappa^2(x,x)$. This supremum exists as $\mathbb{X}$ is compact and $\kappa$ is continuous. The same bound applies to $\widetilde{C}_{XX} - C_{XX}$. For bounding the second term on the right-hand-side of (38), we introduce additional notation. Let $\Pi_{\mathcal{D}}$ be the (linear) projection operator on the closed subspace $\{\varphi(x_i,y_i) : i \in \mathcal{I}\}$ of $\mathcal{H}_\otimes$, where

[2] This should be replaced by: we assume the sparse estimate $\widehat{C}_{XX}$ is positive semi-definite. Such assumption is satisfied when coefficients $\beta \geq 0$.

recall that $\mathcal{I}$ is the set of indices among $\mathcal{M}$ that are present in $\mathcal{D}$. Then, we have

$$\begin{aligned}\left\|\widehat{C}_{XY}-\widetilde{C}_{XY}\right\|_{\text{HS}} &= \left\|\frac{1}{m}\sum_{i=1}^{m}\left(1-\Pi_{\mathcal{D}}\right)\varphi\left(x_i,y_i\right)\right\|_{\mathcal{H}_\otimes}\\ &\leq \sum_{i=1}^{m}\frac{1}{m}\left\|\varphi\left(x_i,y_i\right)-\Pi_{\mathcal{D}}\varphi(x_i,y_i)\right\|_{\mathcal{H}_\otimes}\\ &= \frac{1}{m}\sum_{i\notin\mathcal{I}}\left\|\varphi\left(x_i,y_i\right)-\Pi_{\mathcal{D}}\varphi\left(x_i,y_i\right)\right\|_{\mathcal{H}_\otimes}.\end{aligned} \tag{41}$$

Pythagoras theorem gives

$$\begin{aligned}&\left\|\varphi\left(x_i,y_i\right)-\Pi_{\mathcal{D}}\varphi\left(x_i,y_i\right)\right\|_{\mathcal{H}_\otimes}^2\\ &= \underbrace{\left\|\varphi\left(x_i,y_i\right)\right\|_{\mathcal{H}_\otimes}^2}_{T_1}-\underbrace{\left\|\Pi_{\mathcal{D}}\varphi\left(x_i,y_i\right)\right\|_{\mathcal{H}_\otimes}^2}_{T_2},\end{aligned} \tag{42}$$

where $T_1\leq B$ and we can bound $T_2$ from below as

$$\begin{aligned}\left\|\Pi_{\mathcal{D}}\varphi(x_i,y_i)\right\| &= \max_{\beta}\left\langle\frac{\sum_{j\in\mathcal{I}}\beta_j\varphi(x_j,y_j)}{\left\|\sum_{j\in\mathcal{I}}\beta_j\varphi(x_j,y_j)\right\|},\varphi(x_i,y_i)\right\rangle\\ &= \max_{\beta}\frac{\sum_{j\in\mathcal{I}}\beta_j\kappa_\otimes\Big((x_i,y_i),(x_j,y_j)\Big)}{\left\|\sum_{j\in\mathcal{I}}\beta_j\varphi(x_j,y_j)\right\|}\\ &\geq \max_{q\in\mathcal{I}}\frac{\left|\kappa_\otimes\Big((x_i,y_i),(x_q,y_q)\Big)\right|}{\sqrt{\kappa_\otimes\Big((x_q,y_q),(x_q,y_q)\Big)}}\\ &\geq \gamma\sqrt{\kappa_\otimes\Big((x_i,y_i),(x_i,y_i)\Big)},\end{aligned} \tag{43}$$

where the first inequality results from a specific choice of coefficients. Specifically, it is obtained with $\beta_j=0$ for each $j\in\mathcal{I}$, expect for a single index $q$ with $\beta_q=\pm1$, depending on the sign of $\kappa_\otimes\Big((x_i,y_i),(x_q,y_q)\Big)$. Combining the bounds on $T_1$ and $T_2$ in (42), we get

$$\begin{aligned}\left\|\varphi\left(y_i,x_i\right)-\Pi_{\mathcal{D}}\varphi\left(y_i,x_i\right)\right\|^2 &\leq \kappa_\otimes\left((x_i,y_i),(x_i,y_i)\right)\left(1-\gamma^2\right)\\ &\leq B(1-\gamma^2).\end{aligned} \tag{44}$$

Utilizing this bound in (41), we get

$$\left\|\widehat{C}_{XY}-\widetilde{C}_{XY}\right\|_{\text{HS}}\leq\left(1-\frac{|\mathcal{D}|}{m}\right)\sqrt{B(1-\gamma^2)}. \tag{45}$$

Combining (40) and (45) in (38), we obtain

$$\left\|\widehat{C}_{XY}-C_{XY}\right\|_{\text{HS}}\leq\sqrt{B}\psi(m,\gamma;\delta) \tag{46}$$

with probability $\geq 1-\delta$. The same bound applies to the sparse approximation $\widehat{C}_{XX}$ of $C_{XX}$, using which we obtain our required result as

$$\left\|\widehat{\mathcal{K}}_\varepsilon-\mathcal{K}_\varepsilon\right\|\leq\sqrt{B}\psi(m,\gamma;\delta)\left(\frac{1}{\varepsilon}+\frac{\|C_{XY}\|}{\varepsilon^2}\right). \tag{47}$$

## VI. Computing Eigenfunctions of Sparse Kernel Koopman Operator

The spectra of transfer operators are rich in information: they can be used to decompose modes of a dynamical system, propagate uncertainties, analyze region of attraction of nonlinear dynamical systems, etc. In this section, we show that eigenfunctions of the sparse kernel Koopman operator can be constructed using only Gram matrices over sparse data. To that end, define

$$\Phi_X=\left[\phi(x_1)\dots\phi(x_d)\right],\quad \Phi_Y=\left[\phi\left(y_1\right)\dots\phi(y_d)\right],$$

where $d=|\mathcal{D}|$. Recall from (26) that the regularized sparse kernel Koopman estimator is given by $\widehat{\mathcal{K}}_\varepsilon=\left(\widehat{C}_{XX}+\varepsilon I\right)^{-1}\widehat{C}_{XY}$, defined using the sparse covariance and the cross-covariance operators, that in turn can be written as

$$\begin{aligned}\widehat{C}_{XY}=\Phi_X A_\alpha\Phi_Y^\top,\quad \widehat{C}_{XX}=\Phi_X A_\beta\Phi_X^\top,\\ A_\alpha=\operatorname{diag}\left(\boldsymbol{\alpha}\right),\quad A_\beta=\operatorname{diag}\left(\boldsymbol{\beta}\right),\end{aligned}$$

per (22) and (23). Define the Gram matrices

$$G_{XX}=\Phi_X^\top\Phi_X, G_{YX}=\Phi_Y^\top\Phi_X. \tag{48}$$

Using this notation, rewrite $\widehat{\mathcal{K}}_\varepsilon$ as

$$\begin{aligned}\widehat{\mathcal{K}}_\varepsilon &= \left(\widehat{C}_{XX}+\varepsilon I\right)^{-1}\widehat{C}_{XY}\\ &= \left(\Phi_X A_\beta\Phi_X^\top+\varepsilon I\right)^{-1}\Phi_X A_\alpha\Phi_Y^\top\\ &= \Phi_X\left(A_\beta\Phi_X^\top\Phi_X+\varepsilon I\right)^{-1}A_\alpha\Phi_Y^\top\\ &= \Phi_X\underbrace{\left(A_\beta G_{XX}+\varepsilon I\right)^{-1}A_\alpha}_{\Upsilon}\Phi_Y^\top\\ &= \Phi_X\Upsilon\Phi_Y^\top,\end{aligned} \tag{49}$$

where the third equality follows from the identity

$$(I+PQ)^{-1}P=P(I+QP)^{-1}. \tag{50}$$

Consider the finite dimensional matrix

$$\underbrace{\left(A_\beta G_{XX}+\varepsilon I\right)^{-1}A_\alpha}_{\Upsilon}G_{YX}:=\Upsilon G_{YX} \tag{51}$$

From [5, Proposition 3.1], we get that an operator of the form $\widehat{\mathcal{K}}_\varepsilon=\Phi_X\Upsilon\Phi_Y^\top$ has an eigenvalue $\lambda$ with the corresponding eigenfunction

$$\varphi_\lambda(x)=\Psi(x)\boldsymbol{v},\quad [\Psi(x)]_i=\kappa(x_i,x),\quad i\in\mathcal{I}, \tag{52}$$

if and only if $\boldsymbol{v}$ is a right eigenvector of $\Upsilon G_{YX}$ associated with the same eigenvalue. Such observation enables us to construct eigenfuntions of $\widehat{\mathcal{K}}_\varepsilon$ from finite dimensional Gram matrices $G_{XX}$ and $G_{YX}$.

## VII. A Numerical Example

We empirically evaluate the impact of dictionary sparsification on the spectrum of the empirical kernel Koopman operator for an unforced Duffing oscillator. The dynamical system of the oscillator is described by $\ddot{z} = -\delta\dot{z} - z\left(\beta + \alpha z^2\right)$. We choose $\delta = 0.5$, $\beta = -1$, and $\alpha = 1$. Figure 1 is an illustration of the dynamics of the oscillator; this system exhibits two regions of attraction, corresponding to two equilibrium points $(-1, 0)$ and $(1, 0)$. To compute the eigenfunction of

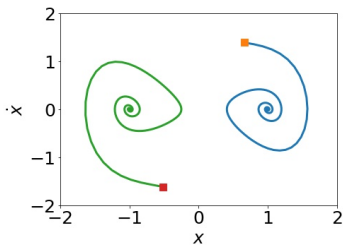
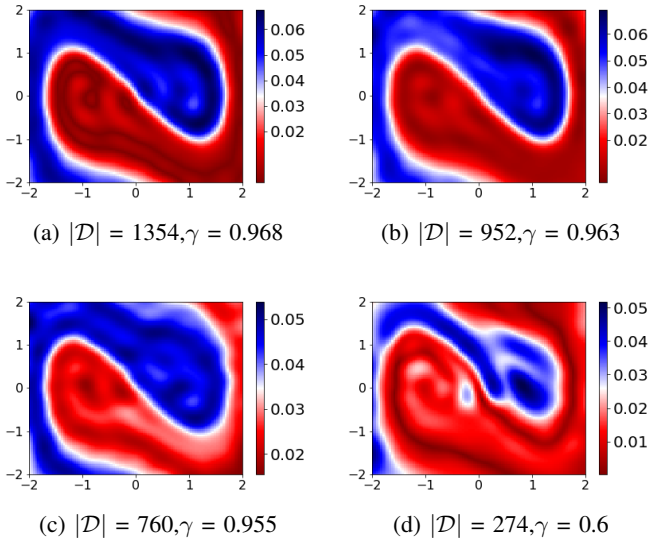



Fig. 1: Two trajectories of the Duffing oscillator that converge to two different equilibrium points.

the sparse kernel Koopman operator, we sampled 1600 initial points $x = [z, \dot{z}]$ that are uniformly distributed over $(z, \dot{z}) \in [-2, 2] \times [-2, 2]$. We then numerically integrate $x$ one step forward with a time interval of $\Delta t = 0.25$ to get $y$. Hence, our dataset consists of 1600 sampled $(x, y)$ pairs. We use a combination of three Gaussian kernels $\kappa(x, y) = \frac{1}{2}\exp\left(-\frac{\|x-y\|^2}{2\times 1.45^2}\right) + \frac{3}{10}\exp\left(-\frac{\|x-y\|^2}{2\times 0.48^2}\right) + \frac{1}{5}\exp\left(-\frac{\|x-y\|^2}{2\times 0.29^2}\right)$ and construct four sparse dictionaries with four different values of $\gamma$. For each dictionary, we plot heat-maps of the leading eigenfunction of the approximate embedded Koopman operator in Figures 2b, 2a, 2c and 2d, using the procedure outlined in Section VI. Upon decreasing $\gamma$, the dictionary becomes more sparse (with less $|\mathcal{D}|$). As shown in Figures 2b, 2a and 2c, the resulting eigenfunctions accurately reveal the distinct regions of attractions. The characterization becomes less sharp with lower values of $\gamma$. In Figure 2d, we plot the leading eigenfuntion obtained with $\gamma = 0.6$. Evidently, it fails to accurately capture the region of attraction. In other words, we lose useful information of the nonlinear dynamics in this case, as we "throw away" too many points with this low a value of $\gamma$.

We remark that we set $\varepsilon = 10^{-10} \times m^{-0.2}$. We construct $\mathcal{D}$ in an online fashion, following the procedure in [22, Section 4.3]. Suppose at time $t + 1$, we have collected a $\gamma$-coherent dictionary $\mathcal{D}_t$ with index set $\mathcal{I}_t$ and are presented with a new sample $(x_{t+1}, y_{t+1})$. If $(x_{t+1}, y_{t+1})$ satisfies the coherence condition (21), we update the dictionary by including the new pair $\mathcal{D}_{t+1} = \mathcal{D}_t \cup \{(x_{t+1}, y_{t+1}\}$. Otherwise, we dismiss the candidate data pair. We remark that our bound in Theorem 1 applies to the online setting as it only requires that the off-diagonal entries of the dictionary ultimately satisfy the coherence condition in (21).




(a) $|\mathcal{D}| = 1354, \gamma = 0.968$ (b) $|\mathcal{D}| = 952, \gamma = 0.963$

(c) $|\mathcal{D}| = 760, \gamma = 0.955$ (d) $|\mathcal{D}| = 274, \gamma = 0.6$

Fig. 2: Plot of the leading eigenfunction of the empirical embedded Koopman operator with coherency-based sparsification of data.

## VIII. Conclusions and Future Directions

In this paper, we present a sparse learning approach for transfer operators that interact with RKHS. We analyze sample complexity for coherence-based sparsification and illustrate its efficacy empirically.

There are a number of interesting directions for future research. First, we want to extend our results to the case where samples are obtained from a continuous trajectory. In such settings, one cannot treat the samples as being independent and identically distributed, and requires a different analysis such as using mixing conditions. Second, we want to explore the use of sparse kernel learning of generators for these operators for continuous-time dynamical systems. Third, we aim to analyze the efficacy of our sparse learning framework to diverse applications, such as model order reduction and uncertainty propagation.

## References


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

[2] Q. Li, F. Dietrich, E. M. Bollt, and I. G. Kevrekidis, "Extended dynamic mode decomposition with dictionary learning: A data-driven adaptive spectral decomposition of the koopman operator," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 27, no. 10, p. 103111, 2017.

[3] M. O. Williams, C. W. Rowley, and I. G. Kevrekidis, "A kernel-based approach to data-driven koopman spectral analysis," *arXiv preprint arXiv:1411.2260*, 2014.

[4] N. Aronszajn, "Theory of reproducing kernels," *Transactions of the American mathematical society*, vol. 68, no. 3, pp. 337–404, 1950.

[5] S. Klus, I. Schuster, and K. Muandet, "Eigendecompositions of transfer operators in reproducing kernel hilbert spaces," *Journal of Nonlinear Science*, vol. 30, no. 1, pp. 283–315, 2020.

[6] K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf, "Kernel mean embedding of distributions: A review and beyond," *arXiv preprint arXiv:1605.09522*, 2016.

[7] K. Fukumizu, F. R. Bach, and A. Gretton, "Statistical consistency of kernel canonical correlation analysis." *Journal of Machine Learning Research*, vol. 8, no. 2, 2007.

[8] K. Fukumizu, L. Song, and A. Gretton, "Kernel bayes' rule: Bayesian inference with positive definite kernels," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 3753–3783, 2013.

[9] M. Wu, B. Schölkopf, G. Bakır, and N. Cristianini, "A direct method for building sparse kernel learning algorithms." *Journal of Machine Learning Research*, vol. 7, no. 4, 2006.

[10] C. Richard, J. C. M. Bermudez, and P. Honeine, "Online prediction of time series data with kernels," *IEEE Transactions on Signal Processing*, vol. 57, no. 3, pp. 1058–1067, 2008.

[11] A. Koppel, G. Warnell, E. Stump, and A. Ribeiro, "Parsimonious online learning with kernels via sparse projections in function space," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 4671–4675.

[12] A. Berlinet and C. Thomas-Agnan, *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

[13] K. Fukumizu, F. R. Bach, and M. I. Jordan, "Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces," *Journal of Machine Learning Research*, vol. 5, no. Jan, pp. 73–99, 2004.

[14] I. Steinwart and A. Christmann, *Support vector machines*. Springer Science & Business Media, 2008.

[15] L. Song, J. Huang, A. Smola, and K. Fukumizu, "Hilbert space embeddings of conditional distributions with applications to dynamical systems," in *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009, pp. 961–968.

[16] Y. Engel, S. Mannor, and R. Meir, "Sparse online greedy support vector regression," in *European Conference on Machine Learning*. Springer, 2002, pp. 84–96.

[17] ——, "The kernel recursive least-squares algorithm," *IEEE Transactions on signal processing*, vol. 52, no. 8, pp. 2275–2285, 2004.

[18] J. Kivinen, A. J. Smola, and R. C. Williamson, "Online learning with kernels," *IEEE transactions on signal processing*, vol. 52, no. 8, pp. 2165–2176, 2004.

[19] Z. Wang, K. Crammer, and S. Vucetic, "Breaking the curse of kernelization: Budgeted stochastic gradient descent for large-scale svm training," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 3103–3131, 2012.

[20] E. C. Cortes and C. Scott, "Sparse approximation of a kernel mean," *IEEE Transactions on Signal Processing*, vol. 65, no. 5, pp. 1310–1323, 2016.

[21] I. Tolstikhin, B. K. Sriperumbudur, and K. Muandet, "Minimax estimation of kernel mean embeddings," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 3002–3048, 2017.

[22] Z. Noumir, P. Honeine, and C. Richard, "Online one-class machines based on the coherence criterion," in *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*. IEEE, 2012, pp. 664–668.