Sparse Learning of Dynamical Systems in RKHS: An Operator-Theoretic Approach

Boya Hou [1] **Sina Sanjari** [1] **Nathan Dahlin** [1] **Subhonmesh Bose** [1] **Umesh Vaidya** [2]

Abstract

Transfer operators provide a rich framework for representing the dynamics of very general, nonlinear dynamical systems. When interacting with reproducing kernel Hilbert spaces (RKHS), descriptions of dynamics often incur prohibitive data storage requirements, motivating dataset *sparsification* as a precursory step to computation. Further, in practice, data is available in the form of trajectories, introducing correlation between samples. In this work, we present a method for sparse learning of transfer operators from β-mixing stochastic processes, in both discrete and continuous time, and provide sample complexity analysis extending existing theoretical guarantees for learning from non-sparse, i.i.d. data. In addressing continuous-time settings, we develop precise descriptions using covariance-type operators for the infinitesimal generator that aids in the sample complexity analysis. We empirically illustrate the efficacy of our sparse embedding approach through deterministic and stochastic nonlinear system examples.

1. Introduction

Transfer operators such as the Koopman and the Perron-Frobenius (PF) operators are central to global analysis of complex dynamical systems across a variety of fields, including biology, engineering, finance, and physics. In contrast to direct finite-dimensional, nonlinear state-space descriptions, operator approaches offer infinite-dimensional, but *linear* system models. The spectra of such operators can be utilized to characterize basins of attraction, perform model reduction, propagate uncertainties and analyze global stability of the dynamics among other application uses. Taking a biological context, for instance, protein folding conformations can be understood in terms of metastable sets, which in turn may be estimated using the eigenvalues and eigenfunctions of transfer operators describing underlying molecular dynamics (Klus et al., 2020a).

Practically speaking, approximations to transfer operators can be computed from data. Of the existing parametric methods available, extended dynamic mode decomposition (EDMD) (Williams et al., 2015), and its continuous-time analog Mauroy & Goncalves (2019); Klus et al. (2020b); Nüske et al. (2023) are perhaps the most frequently used. Given the difficulty of selecting a proper set of basis functions in such techniques, non-parametric approximation methods employing operator embeddings within RKHS were proposed in (Williams et al., 2014) and (Klus et al., 2020a) for discrete and continuous time systems, respectively. While such methods typically enjoy sample complexity bounds independent of the underlying system state dimension, they suffer from other drawbacks. In particular, as kernel-based methods automatically produce a set of basis functions from data, the resulting system descriptions grow with input dataset size, making scalability a key hurdle to their usage (Lever et al., 2016).

Sparsification, the process of discarding selected input data found to be redundant in some sense, is a common approach to improving scalability of kernel methods (Engel et al., 2002; Wu et al., 2006; Richard et al., 2008; Koppel et al., 2017). In this paper, we study the sample complexity impacts of coherency-based (Richard et al., 2008) *sparse* learning on transfer operators interacting with RKHS. To the best of our knowledge, for discrete-time settings such as Markov chains, existing analyses in this vein have been limited to the following cases: (i) embeddings are estimated from sparsified i.i.d. input/output data (Hou et al., 2021), and thus are not applicable to the case where only a collection of trajectories is available (ii) the process is mixing but stationary (Mollenhauer et al., 2020; Kostic et al., 2022), and no sparsification is considered. Here, under the assumption that system dynamics can be described by a non-stationary, β-mixing process which converges to a unique stationary

[1]Department of Electrical and Computer Engineering at the University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA [2]Department of Mechanical Engineering, Clemson University, Clemson SC, 29634, USA. Correspondence to: Boya Hou <boyahou2@illinois.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

distribution, we establish sample complexity bounds for transfer operator learning, given correlated, non-stationary, and sparsified data.

In many applications, it is desirable to study continuous-time system models. Operator theoretic analysis of such systems centers around the estimation of the infinitesimal *generator* of transfer operator families, parameterized in time. Prior work has connected such generators to RKHS, and offered effective estimation algorithms under minimal assumptions (Klus et al., 2020a; Rosenfeld et al., 2019). Still, key theoretical properties of embedded generators, including their domain and the continuity of their associated operator families have yet not been rigorously established, a gap that this work bridges. Further, embedding generators within an RKHS requires that partial derivatives of RKHS elements can be represented within the same RKHS, a result that heretofore has only been established in cases where the underlying state space is compact (Zhou, 2008). As we target SDE-driven dynamics where the system domain is not compact, we develop an alternative sufficient condition for closedness under partial differentiation based upon boundedness and decay of the partial derivatives of the kernel.

The representation we develop for embedded generators in terms of covariance-type operators allows for application of our discrete-time sample complexity analysis methodology to continuous-time settings. We prove sample complexity bounds for learning embeddings of generators corresponding to non-stationary, β-mixing processes.

Our key contributions are as follows: (a) We provide sample complexity bounds for learning transfer operators from sparsified data produced by non-stationary β-mixing processes. (b) We define generators of such operator families for continuous-time systems, and characterize their domains among other properties. For this purpose, we study partial derivatives of functions in RKHS over non-compact spaces. (c) We characterize these generators in terms of covariance–type operators and provide sample complexity guarantees for learning them in RKHS.

Section 2 serves as a prerequisite for learning dynamical systems using RKHS. For the discrete-time case, a sparse learning algorithm based on both i.i.d. samples and trajectories is proposed in Section 3 and 4, followed by theoretical analysis. Sections 5 and 6 are devoted to defining and characterizing generators of transfer operator families in continuous-time settings. A data driven algorithm with sample complexity guarantees is presented in Section 7.

2. RKHS and Conditional Mean Embedding

We begin by formally defining an RKHS. See Muandet et al. (2016) for an introduction. Let $\mathbb{X}$ be a 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 κ – 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, ϕ is called the feature map for kernel κ. The inner product satisfies the reproducing property, given by $\langle \phi(x), f \rangle = f(x)$ for all $x \in \mathbb{X}$ and $f \in \mathcal{H}$. The norm associated with $\mathcal{H}$ is defined as $\|f\|_{\mathcal{H}} = \sqrt{\langle f, f \rangle}$ for $f \in \mathcal{H}$. Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with Borel σ-algebra $\mathcal{F}$ and a probability measure $\mathbb{P}$. Let $X : \Omega \to \mathbb{X}$ be a random variable with distribution $\mathbb{P}_X$. The *kernel mean embedding* (KME) of $\mathbb{P}_X$ in $\mathcal{H}$ is the Bochner integral

$$\mu_{\mathbb{P}_X} := \mathbb{E}_X \left[\kappa(X, \cdot) \right] \quad \text{for } X \sim \mathbb{P}_X. \tag{1}$$

Under measurability and boundedness assumptions on κ (see Muandet et al. (2016, Lemma 3.1)), $\mu_{\mathbb{P}_X} \in \mathcal{H}$.

Suppose that $\mathbb{P}(X, Y)$ denotes a joint distribution over $\mathbb{X} \times \mathbb{X}$, then $\mathbb{P}(X, Y)$ can be embedded in the tensor product space $\mathcal{H}_\otimes := \mathcal{H} \otimes \mathcal{H}$, per Berlinet & Thomas-Agnan (2011), as

$$C_{XY} := \mathbb{E}_{XY}[\phi(X) \otimes \phi(Y)] = \mu_{\mathbb{P}_{XY}}, \tag{2}$$

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

$$\kappa_\otimes((x_1, y_1), (x_2, y_2)) = \kappa(x_1, x_2)\, \kappa(y_1, y_2), \tag{3}$$

for x_1, x_2, y_1, y_2 in $\mathbb{X}$. Its joint feature map is $\varphi(x_i, y_i) = \kappa(x_i, \cdot)\, \kappa(y_i, \cdot)$. Let $\mathrm{HS}(\mathcal{H})$ be the Hilbert space of Hilbert-Schmidt (HS) operators from $\mathcal{H}$ to $\mathcal{H}$, endowed with the norm $\|A\|_{\mathrm{HS}}^2 = \sum_{i \in \mathbb{N}} \|Ae_i\|_{\mathcal{H}}^2$ for $A \in \mathrm{HS}(\mathcal{H})$, where $\{e_i\}_{i \in \mathbb{N}}$ is an orthonormal basis of $\mathcal{H}$. In (2), we identify C_{XY} as an element in the tensor product space. Since $\mathcal{H} \otimes \mathcal{H}$ is isometrically isomorphic to $\mathrm{HS}(\mathcal{H})$ (Aubin, 2011), it can also be viewed as an HS operator $C_{XY} \in \mathrm{HS}(\mathcal{H})$ that satisfies

$$\mathbb{E}_{XY}[f(X)g(Y)] = \langle C_{XY} g, f \rangle, \quad \forall f, g \in \mathcal{H}. \tag{4}$$

C_{XY} is called the (uncentered) *cross-covariance* operator. Also, define the (uncentered) covariance operator as

$$C_{XX} := \mathbb{E}_X[\phi(X) \otimes \phi(X)], \tag{5}$$

which can be viewed as the embedding of the marginal distribution $\mathbb{P}_X$ in $\mathcal{H}_\otimes$. Throughout this paper, we make the following standing assumption.

Assumption 1. *The kernel κ is continuous and bounded as $\sup_{x \in \mathbb{X}} \kappa(x, x) \leq B_\kappa < \infty$ for some $B_\kappa \in \mathbb{R}$.*

According to Steinwart & Christmann (2008, Chapter 4), the feature map ϕ and $\kappa : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ are measurable under Assumption 1.

Let $\mathbb{P}_{Y|x}$ denote the conditional distribution of Y, given $X = x \in \mathbb{X}$. The $\mathcal{H}$-embedding of $\mathbb{P}_{Y|x}$ is

$$\mu_{\mathbb{P}_{Y|x}} := \mathbb{E}_{Y|x}[\phi(Y)|X = x] \qquad \forall x \in \mathbb{X}. \tag{6}$$

Per Song et al. (2009), the *conditional mean embedding* (CME) operator $\mathcal{U}_{Y|X} : \mathcal{H} \to \mathcal{H}$ is a linear operator that satisfies $\mu_{\mathbb{P}_{Y|x}} = \mathcal{U}_{Y|X}\phi(x)$. If C_{XX} is injective and $\mathbb{E}_{Y|x}[f(Y)|X = x] \in \mathcal{H}$ for all $f \in \mathcal{H}$ and $x \in \mathbb{X}$, then $\mathcal{U}_{Y|X} = C_{YX}C_{XX}^{\dagger}$. For technical reasons, we consider its regularized version,

$$\mathcal{U}_{\varepsilon} = C_{YX}\left(C_{XX} + \varepsilon\,\mathrm{id}\right)^{-1}, \tag{7}$$

for $\varepsilon > 0$, where id is the identity operator.

3. Transfer Operators via CME Operator

Let $\mathbb{T}$ be the set of nonnegative integers and $\{X_t\}_{t\in\mathbb{T}}$ be a $\mathbb{R}^n$-valued time-homogeneous Markov process defined on a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\in\mathbb{T}}, \mathbb{P})$, where X_t is the $\mathcal{F}_t$-adapted state of the system at time t. Such a stochastic dynamical system can be described by the transition kernel density p as $\mathbb{P}\{X_{t+1} \in \mathbb{A}|X_t = x\} = \int_{\mathbb{A}} p(y|x)\mathrm{d}y$ for measurable $\mathbb{A} \subseteq \mathbb{R}^n$. If f is a probability density over $\mathbb{R}^n$, then the Perron–Frobenius (PF) operator $\mathcal{P} : L^1(\mathbb{R}^n) \to L^1(\mathbb{R}^n)$ propagates $f \in L^1(\mathbb{R}^n)$ as

$$\left(\mathcal{P}f\right)(y) = \int p(y|x)f(x)\mathrm{d}x. \tag{8}$$

If f is a scalar function of $\mathbb{R}^n$, then the Koopman operator $\mathcal{K} : L^{\infty}(\mathbb{R}^n) \to L^{\infty}(\mathbb{R}^n)$ acts on $f \in L^{\infty}(\mathbb{R}^n)$ as

$$\left(\mathcal{K}f\right)(x) = \int p(y|x)f(y)\mathrm{d}y. \tag{9}$$

These transfer operators are infinite-dimensional but linear. They are related to CME as follows. When interacting with RKHS, $\mathcal{P}$ propagates the embedded distribution of states through the system dynamics. Now, let X^+ be the system state at the next time-step starting from X. We have that $\mathcal{U}_{X^+|X} : \mu_{\mathbb{P}_X} \mapsto \mu_{\mathbb{P}_{X^+}}$ and it satisfies $\mu_{\mathbb{P}_{X^+}} = \mathcal{U}_{X^+|X}\mu_{\mathbb{P}_X}$ per Song et al. (2009); Hou et al. (2021). Hence, we identify $\mathcal{P}$ as the CME operator $\mathcal{U}_{X^+|X} = C_{X^+X}C_{XX}^{\dagger}$ and its regularized variant as $\mathcal{P}_{\varepsilon} := \mathcal{U}_{\varepsilon}$. Furthermore, the Koopman operator satisfies

$$\langle \mathcal{K}f, \phi(x)\rangle = \left\langle f, \mu_{X^+|X}\right\rangle = \left\langle f, \mathcal{U}_{X^+|X}\phi(x)\right\rangle \tag{10}$$

for all $f \in \mathcal{H}$. Thus, $\mathcal{K}$ is the adjoint of $\mathcal{U}_{X^+|X} = \mathcal{P}$, given by $\mathcal{K} := C_{XX}^{\dagger}C_{XX^+}$. In this paper, we learn regularized variants of $\mathcal{P}$ and $\mathcal{K}$ from data. While we report sample complexity results for learning variants of $\mathcal{P}$, they also apply to the variants of $\mathcal{K}$.

Covariance operators in (7) can be estimated from data via sample average approximation given by (2). However, such description grows with the size of the dataset as kernel functions centered around each data point are added to the empirical operator description; see Engel et al. (2002); Kivinen et al. (2004); Koppel et al. (2017) for discussions. This work extends the framework proposed in Hou et al. (2021) that reduces the dictionary $\mathcal{D}$ to $\mathcal{D}_{\gamma}$ using the notion of *coherency* from Richard et al. (2008)). For a given dataset $\mathcal{D}$ of M points $(x_1, x_1^+), \ldots, (x_M, x_M^+)$, we construct $\mathcal{D}_{\gamma}$ by identifying a subset that satisfies

$$\left|\kappa(x_i^{\star}, x_j^{\star})\right| \leq \sqrt{\gamma\kappa(x_i^{\star}, x_i^{\star})\kappa(x_j^{\star}, x_j^{\star})}, \tag{11}$$

for each i, j, where $(x_i^{\star}, x_j^{\star})$ is either (x_i, x_j) or (x_i^+, x_j^+), and $(x_i, x_i^+), (x_j, x_j^+)$ are in $\mathcal{D}_{\gamma}$. One can construct such a $\mathcal{D}_{\gamma}$ using the Gram matrix with all elements in $\mathcal{D}$; see (Hou et al., 2021) for details. Let $\mathcal{I}$ be the indices among $1, \ldots, M$ for which (x_i, x_i^+) are in $\mathcal{D}_{\gamma}$. Then, the sparse covariance operator estimates are

$$\widehat{C}_{X^+X} = \sum_{i\in\mathcal{I}} \alpha_i \varphi\left(x_i^+, x_i\right), \widehat{C}_{XX} = \sum_{i\in\mathcal{I}} \beta_i \varphi\left(x_i, x_i\right), \tag{12}$$

where α (and similarly, β) is defined via $\alpha = G^{-1}g$, and $G \in \mathbb{R}^{|\mathcal{D}_{\gamma}|\times|\mathcal{D}_{\gamma}|}$, $g \in \mathbb{R}^{|\mathcal{D}_{\gamma}|}$ are $G_{i,j} = \kappa_{\otimes}((x_i^+, x_i), (x_j^+, x_j))$ and $g_j = \frac{1}{M}\sum_{r=1}^{M} G_{r,j}$ for each i and j in $\mathcal{I}$. The compressed covariance operators then become $\widehat{C}_{\star}$, $\star = \{X^+X, XX\}$. The sparse PF estimator is then defined as $\widehat{\mathcal{P}}_{\varepsilon} := \widehat{C}_{X^+X}\left(\widehat{C}_{XX} + \varepsilon\,\mathrm{id}\right)^{-1}$. The following result corrects a minor error in (Hou et al., 2021), and is proven in Appendix A. We use the notation $\|\cdot\|$ to denote operator norm, and define

$$\Xi(\nu) := 1 + \sqrt{2\log(1/\nu)}, \quad \nu \in \mathbb{R}.$$

Theorem 1. *Let $\mathcal{D} = \left((x_1, x_1^+), \cdots, (x_M, x_M^+)\right)$ be a dataset with M i.i.d. samples. Then, under Assumption 1, $\left\|\widehat{\mathcal{P}}_{\varepsilon} - \mathcal{P}_{\varepsilon}\right\| \leq B_{\kappa}\psi_1(M, \gamma; \delta/2)\mathcal{O}(\varepsilon^{-2})$ [1] with probability at least $1 - \delta$ for $\delta \in (0, 1)$, if the sparse estimate $\widehat{C}_{XX}$ is positive semi-definite [2], where*

$$\psi_1(M, \gamma; \delta) := \frac{1}{\sqrt{M}}\Xi(\delta) + \left(1 - \frac{|\mathcal{D}_{\gamma}|}{M}\right)\sqrt{1 - \gamma^2}. \tag{13}$$

[1] It was brought to our attention that an alternate analytical framework presented in (Li et al., 2022), which in turn relies on (Steinwart & Scovel, 2012; Fischer & Steinwart, 2020), holds promise to improve the $\mathcal{O}(\varepsilon^{-2})$ dependency on the regularization parameter ε.

[2] Such an assumption is satisfied when coefficients $\beta \geq 0$. An alternative estimate which satisfies the positive semi-definiteness of $\widehat{C}_{XX}$ is by using uniform weights,i.e., $\alpha = \beta = \frac{1}{|\mathcal{I}|}\mathbf{1}$. This will introduce an additional error term in (13) depending on γ, B_{κ}, M, which encodes the effect of weights adjustments.

Decreasing the coherency parameter γ leads to a less coherent dictionary and a smaller $\mathcal{D}_\gamma$. However, this sparsity comes at the cost of approximation accuracy. One cannot avoid this cost with more data if $|\mathcal{D}_\gamma|$ saturates—a scenario that arises when the dynamics evolves over a compact domain as the next result (proven in Appendix B) reveals.

Theorem 2. *Suppose $\{X_t\}_{t\in\mathbb{T}}$ takes values on a compact subset $\mathbb{K}\subset\mathbb{R}^n$ and $B'_\kappa \le \kappa(x,x) \le B_\kappa$ for all $x\in\mathbb{K}$. If $B'_\kappa \ge \sqrt{\gamma}B_\kappa$, then for any possible sequence $(x_i, x_i^+)\in \mathbb{K}\times\mathbb{K}$, any γ-coherent subset $\mathcal{D}_\gamma$ of the sequence satisfies $|\mathcal{D}_\gamma| \le C_\mathbb{K}/(1-\gamma)^n$ for some value $C_\mathbb{K}$ that depends on $\mathbb{K}$ and κ.*

A kernel κ is translation invariant if $\kappa(x,y)$ can be written as a function of $x-y$. Such kernels, including the Gaussian kernels considered in Section 9, satisfy $B'_\kappa = B_\kappa$, meaning that $|\mathcal{D}_\gamma|$ is bounded for all $0<\gamma<1$.

4. Sparse PF Operator Learning from Trajectories

Samples of a dynamical system are often collected along trajectories, i.e., non-i.i.d or time-correlated data. In this section, we develop sample complexity bounds for PF operator estimation based on data collected from trajectories of discrete-time systems. In our setting, dependencies among samples weaken over time, and samples converge to a unique stationary distribution Π in a suitable sense. Specifically, we assume that the data is drawn from a β-mixing stochastic process, defined below. See Yu (1994); Mohri & Rostamizadeh (2008); Vidyasagar (2013) for other applications.

Definition 1. *Agarwal & Duchi (2012, Definition II.1) β-Mixing. Let $\{X_t\}_{t\in\mathbb{T}}$ be a stochastic process on a filtered probability space $(\Omega, \{\mathcal{F}_t\}_{t\in\mathbb{T}}, \mathbb{P})$ where X_t is $\mathcal{F}_t$-adapted and takes values in $\mathbb{X}=\mathbb{R}^n$. Let $P_{t+s}(\cdot \mid \mathcal{F}_t)$ be a version of the conditional distribution of X_{t+s} given $\mathcal{F}_t$. Assume that Π defines the unique stationary distribution of the stochastic process over $\mathbb{R}^n$. Then, the β-coefficients of $\{X_t\}_{t\in\mathbb{T}}$ are defined as*

$$\beta(s) := \sup_t \mathbb{E}\left\|P_{t+s}(\cdot \mid \mathcal{F}_t) - \Pi\right\|_{TV}, \tag{14}$$

where $\|\cdot\|_{TV}$ is the total variation distance. A process $\{X_t\}_{t\in\mathbb{T}}$ is said to be β-mixing, if $\beta(s)\to 0$ as $s\to\infty$.

In addition, we make the following assumption which requires $P_{t+s}(\cdot \mid \mathcal{F}_t)$ and the stationary density Π to span the state space.

Assumption 2. *$P_{t+s}(\cdot \mid \mathcal{F}_t)$ and Π are absolutely continuous with respect to the Lebesgue measure on $\mathbb{X}$ for all $t,s\in\mathbb{T}$.*

Before stating our results for the CME operator, we first present a theorem on learning KME with trajectories. In particular, sample complexity guarantees for learning KME (1) from i.i.d. data are studied in Smola et al. (2007); Gretton et al. (2012); Lopez-Paz et al. (2015); Tolstikhin et al. (2017). We extend these results to β-mixing stochastic processes.

Theorem 3. *Let $\{X_t\}_{t\in\mathbb{T}}$ be a β-mixing stochastic process defined on a filtered probability space $(\Omega,\mathcal{F},\{\mathcal{F}_t\}_{t\in\mathbb{T}},\mathbb{P})$ where X_t is $\mathcal{F}_t$-adapted and takes values in $\mathbb{X}=\mathbb{R}^n$, with a stationary distribution Π. Starting from X_0 sampled according to an initial distribution P_{X_0} and evolving through the system dynamics, let $X^s(m) := \left(X_{(1)},\ldots,X_{(m)}\right)$ be an m-length sequence of states, sampled s time-points apart. Suppose further that Assumptions 1 and 2 hold. Then, the empirical KME estimate $\widetilde{\mu}(X^s(m)) = \frac{1}{m}\sum_{j=1}^m \kappa(X_{(j)},\cdot)$ and the embedding μ of Π satisfy*

$$\left\|\widetilde{\mu}(X^s(m)) - \mu\right\|_{\mathcal{H}} \le \sqrt{\frac{B_\kappa}{m}}\,\Xi\left(\delta - m\beta(s)\right) \tag{15}$$

with probability at least $1-\delta$ for any $m\beta(s)<\delta<1$.

According to Tolstikhin et al. (2017, Proposition A.1), KME learning from i.i.d. data achieves $\mathcal{O}(m^{-1/2})$-consistency with m samples. The same learning rate is thus preserved in the β-mixing case when sub-samples are separated by s steps. The spacing s controls the Markovian dependency between subsampled data points. Increasing s reduces $\beta(s)$ and tightens the bound. This tightening, however, comes at the cost of discarding more samples.

One can utilize the techniques from Section 3 to construct a sparse PF operator from a sub-sampled β-mixing dataset. Theorem 3 applied to estimates in $\mathcal{H}_\otimes$ yields $\left\|\widetilde{C}_{X^+X} - C_{X^+X}\right\|_{\text{HS}} \le B_\kappa^2\,\Xi\left(\delta - m\beta(s)\right)/\sqrt{m}$ with probability at least $1-\delta$ for any $m\beta(s)<\delta<1$. The above observation then yields the following result; the proof is similar to that of Theorem 1, and is omitted.

Theorem 4. *Under the same assumptions as Theorem 3, $\left\|\widehat{\mathcal{P}}_\varepsilon - \mathcal{P}_\varepsilon\right\| \le B_\kappa\psi_2(m,s,\gamma;\delta/2)\mathcal{O}(\varepsilon^{-2})$ for $m\beta(s)<\delta<1$, with probability at least $1-\delta$, where $\psi_2(m,s,\gamma;\delta) := \left(1-\frac{|\mathcal{D}_\gamma|}{m}\right)\sqrt{1-\gamma^2} + \frac{1}{\sqrt{m}}\Xi\left(\delta - m\beta(s)\right)$.*

5. PF Semigroup in Continuous Time

For a diffusion process X_t (a time-homogeneous Markov process in continuous time with almost surely continuous sample paths), one can associate a PF operator *family*, parameterized by time. Let $\mathbb{T} := [0,\infty)$. Consider a filtered probability space $(\Omega,\mathcal{F},\{\mathcal{F}_t\}_{t\in\mathbb{T}},\mathbb{P})$, where X_t is $\mathcal{F}_t$-adapted and takes values in $\mathbb{X}=\mathbb{R}^n$. The embedded PF operator $\mathcal{P}_t$ is defined via a Bochner conditional expectation,

$$\mathcal{P}_t\kappa(x,\cdot) = \mathbb{E}[\kappa(X_t,\cdot)|X_0=x]. \tag{16}$$

Let P be the transition function such that for fixed x, $P(t,x;\mathbb{A}) := \mathbb{P}\{X_t \in \mathbb{A} | X_0 = x\}$. The PF operator family $\{\mathcal{P}_t\}_{t\in\mathbb{T}}$ then defines a semigroup because $\mathcal{P}_0 = \mathrm{id}$, and

$$\begin{aligned}\mathcal{P}_{t+s}\kappa(x,\cdot) &= \int \kappa(y,\cdot) P(t+s,x;\mathrm{d}y) \\ &\overset{(a)}{=} \int\int \kappa(y,\cdot) P(s,z;\mathrm{d}y) P(t,x;\mathrm{d}z) \\ &= \int \mathcal{P}_s\kappa(z,\cdot) P(t,x;\mathrm{d}z) \\ &= \mathcal{P}_t \circ \mathcal{P}_s \kappa(x,\cdot),\end{aligned} \tag{17}$$

where Chapman-Kolmogorov equation implies (a). We now study properties of the semigroup of operators $\{\mathcal{P}_t\}_{t\in\mathbb{T}}$. A semigroup is strongly continuous if $\lim_{t\to 0} \|\mathcal{P}_t f - f\|_{\mathcal{H}} = 0$ for all $f \in \mathcal{H}$, and uniformly continuous, if $\lim_{t\to 0} \|\mathcal{P}_t - \mathcal{I}\| = 0$, where $\|\cdot\|$ is the operator norm. Furthermore, define its infinitesimal generator $\mathcal{A}$ as the operator that satisfies

$$\mathcal{A}f := \lim_{t\to 0} \frac{1}{t} \left(\mathcal{P}_t f - f\right), \tag{18}$$

for $f \in \mathcal{H}$. The limit in $\mathcal{H}$-norm exists over a subspace $\mathbb{D}(\mathcal{A})$ of $\mathcal{H}$. Next, we study the properties of $\{\mathcal{P}_t\}_{t\in\mathbb{T}}$ and $\mathcal{A}$. Let C_0 be the covariance operator at time $t = 0$ and C_t be the cross-covariance with time-lag t, i.e., $C_0 := C_{X_0X_0}$ and $C_t := C_{X_tX_0}$. We make the following assumption.

Assumption 3. *$\mathbb{E}[f(X_t)|X = \cdot] \in \mathcal{H}$ for all $f \in \mathcal{H}$, $\forall t \in \mathbb{T}$; and C_0 is invertible.*

In essence, the first part of this assumption imposes closedness of the RKHS $\mathcal{H}$ with respect to the evolution of functions and probability densities through the system dynamics. Such a closedness assumption is not new to Koopman operator-based analysis, e.g.,Yeung et al. (2019); Nandanoori et al. (2020). Nevertheless, it is restrictive and is often violated as noted in (Park & Muandet, 2020; Klebanov et al., 2020). While our proofs require this strong assumption, it may be possible to relax it in light of recent developments in (Li et al., 2022; Kostic et al., 2022). We leave exploration of such a relaxation for future efforts. The proof of the next result is provided in Appendix D.

Theorem 5. *Under Assumptions 1 and 3, $\{\mathcal{P}_t\}_{t\in\mathbb{T}}$ is uniformly continuous in $\mathcal{H}$. Moreover, its generator $\mathcal{A}$ is a bounded linear operator with $\mathbb{D}(\mathcal{A}) = \mathcal{H}$.*

Previous analysis in Klus et al. (2020a) and Rosenfeld et al. (2019) consider generators of transfer operator families. Continuity properties of the operator family and the domain of the generator were not rigorously established to our knowledge–a gap that Theorem 5 bridges.

The Koopman operator family can be defined as the adjoint of the PF operator family, i.e., $\mathcal{K}_t = \mathcal{P}_t^\star$ and satisfies $\langle f, \mathcal{K}_t g\rangle = \langle \mathcal{P}_t f, g\rangle$. Also, $\{\mathcal{K}_t\}_{t\in\mathbb{T}}$ defines a semigroup that satisfies

$$\langle f, (\mathcal{K}_t g - g)/t\rangle = \langle (\mathcal{P}_t f - f)/t, g\rangle \tag{19}$$

for all $f, g \in \mathcal{H}$ and $t > 0$. Taking $t \to 0$ allows us to write the right hand side of (19) as $\langle \mathcal{A}f, g\rangle$, and define the generator of $\{\mathcal{K}_t\}_{t\in\mathbb{T}}$ as $\mathcal{A}^\star$, the adjoint of $\mathcal{A}$. Next, we establish $\mathbb{D}(\mathcal{A}^\star)$; see Appendix E for a proof.

Corollary 1. *Suppose Assumptions 1 and 3 hold. Then, $\mathcal{A}^\star$ is a bounded linear operator with $\mathbb{D}(\mathcal{A}^\star) = \mathcal{H}$.*

6. PF Generator for Stochastic Differential Equations via Covariance-Type Operators

Under Assumption 3, $\mathcal{P}_t = C_t C_0^{-1}$. We now develop an analogous expression for $\mathcal{A}$ in terms of covariance-type operators. To that end, notice that

$$\mathcal{A}\mu = \lim_{t\to 0} \frac{\mathcal{P}_t - \mathrm{id}}{t}\mu = \lim_{t\to 0} \frac{C_t - C_0}{t} C_0^{-1}\mu, \tag{20}$$

where the limit is taken in the $\mathcal{H}$-norm. When $\lim_{t\to 0}(C_t - C_0)/t$ exists, we identify the limit as the operator ∂C_0. In the following, we identify ∂C_0 as an element in the tensor product RKHS. For all $f, g \in \mathcal{H}$, we have

$$\begin{aligned}\langle f, \partial C_0 g\rangle &= \lim_{t\to 0} \frac{1}{t} \left(\langle f, C_t g\rangle - \langle f, C_0 g\rangle\right) \\ &= \lim_{t\to 0} \frac{1}{t} \left(\mathbb{E}[f(X_t)g(X_0)] - \mathbb{E}[f(X_0)g(X_0)]\right) \\ &= \lim_{t\to 0} \frac{1}{t} \left(\mathbb{E}\Big[\big(\mathbb{E}[f(X_t)|X_0] - f(X_0)\big)g(X_0)\Big]\right).\end{aligned} \tag{21}$$

The above derivation uses the definition of C_t, C_0 and the tower property. We simplify the above expression using the Koopman generator $\mathcal{A}^\star$ of $\{X_t\}_{t\in\mathbb{T}}$, characterized by

$$(\mathcal{A}^\star f)(x) := \lim_{t\to 0} \frac{1}{t} \left(\mathbb{E}[f(X_t) - f(X_0)|X_0 = x]\right) \tag{22}$$

for $f \in \mathcal{H}$. For any $\nu > 0$, there is a sufficiently small t_ν, such that for $0 \le t \le t_\nu$, Assumptions 1 and 3 yield

$$\begin{aligned}\left|\frac{1}{t}\left(\mathbb{E}[f(X_t)|x_0] - f(x_0)\right)\right| &\le \nu + |(\mathcal{A}^\star f)(x_0)| \\ &= \nu + |\langle \mathcal{A}^\star f, \kappa(x_0,\cdot)\rangle_{\mathcal{H}}| \\ &\le \nu + \|\mathcal{A}^\star\| \, \|f\|_{\mathcal{H}} \, B_\kappa.\end{aligned}$$

Thus, the dominated convergence theorem allows us to infer from (21) and (22) that

$$\langle f, \partial C_0 g\rangle = \mathbb{E}\left[\mathcal{A}^\star f(X_0) g(X_0)\right]. \tag{23}$$

We now restrict attention to dynamics described by the following stochastic differential equation (SDE),

$$\mathrm{d}X_t = b(X_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}B_t, \quad X_0 = x, \tag{24}$$

where B_t is an n-dimensional Brownian motion. The drift and diffusion functions $b : \mathbb{R}^n \to \mathbb{R}^n$ and $\sigma : \mathbb{R}^n \to \mathbb{R}^{n\times n}$, are assumed Lipschitz continuous, and bounded as

$$\|b_j\|_\infty \leq \overline{S},\ \|a_{i,j}\|_\infty \leq \overline{S}, \quad i,j = 1,\cdots,n. \tag{25}$$

The Koopman generator for such a system, according to Schilling (2021, Theorem 19.9), is

$$\big(\mathcal{A}^\star g\big)(x) = \sum_{i=1}^n b_i(x)\frac{\partial g(x)}{\partial x_i} + \frac{1}{2}\sum_{i,j=1}^n a_{ij}(x)\frac{\partial^2 g(x)}{\partial x_i \partial x_j}, \tag{26}$$

for any $g \in \mathcal{H}$, where $a(x) := \sigma(x)\sigma(x)^\top$. The domain of this operator is $\mathcal{H}$, when a and b are such that Assumption 3 holds. To relate ∂C_0 to $\mathcal{A}^\star$ as defined in (26), we need machinery to describe partial derivatives of functions in $\mathcal{H}$. In Section 6.1, we present these preliminaries and return to defining ∂C_0 in terms of $\mathcal{A}^\star$ in Section 6.2.

6.1. Partial Derivative of Kernel Functions

Representability of partial derivatives of functions in RKHS within the same RKHS has been studied in Zhou (2008), where the analysis considers RKHS over compact domains. We study dynamics driven by an SDE over a non-compact domain, necessitating technical extensions to the analysis in Zhou (2008). We relax compactness and instead consider bounded kernels whose partial derivatives vanish at infinity.

The partial derivatives of κ are denoted by

$$D^{(\alpha,\beta)}\kappa(x,y) = \frac{\partial^{|\alpha|}\partial^{|\beta|}}{\partial x^\alpha \partial y^\beta}\kappa\left(x^1,\ldots,x^n,y^1,\ldots,y^n\right),$$

for $\alpha := (\alpha_1,\cdots,\alpha_n)$ and $\beta := (\beta_1,\cdots,\beta_n)$. Also, let $D^\alpha\kappa(x,\cdot) := D^{(\alpha,0)}\kappa(x,\cdot)$. Define C_∞ to be the family of all continuous functions $f : \mathbb{R}^n\times\mathbb{R}^n \to \mathbb{R}$ such that for each $\epsilon > 0$, there is a compact set $\mathbb{K}_\epsilon$ for which $|f(x,y)| < \epsilon$ for all $(x,y) \in \mathbb{R}^n \times \mathbb{R}^n \setminus \mathbb{K}_\epsilon$. Denote $|\alpha| = \sum_{j=1}^n \alpha_j$. With this notation, we make the following assumption about κ.

Assumption 4. $\left\|D^{(\alpha,\beta)}\kappa\right\|_\infty \leq B_\kappa$ *and* $D^{(\alpha,\beta)}\kappa \in C_\infty$ *for all* $|\alpha+\beta| < 2s$ *for some positive integer* s.

In the following result, we establish that the partial derivative of κ is also an element of $\mathcal{H}$ and satisfies a partial derivative reproducing property. See Appendix F for proof.

Theorem 6. *Let* $I_s := \{\alpha \in \mathbb{N}^n : |\alpha| \leqslant s\}$. *If* κ *satisfies Assumption 4, then for any* $x \in \mathbb{R}^n$, $f \in \mathcal{H}$ *and* $\alpha \in I_s$, *we have* $D^\alpha\kappa(x,\cdot) \in \mathcal{H}$, *and* $\left(D^\alpha f\right)(x) = \langle D^\alpha\kappa(x,\cdot), f\rangle_{\mathcal{H}}$.

6.2. Writing ∂C_0 in Tensor Product Hilbert Space

Define the second order differential operator

$$\mathrm{d}^{(2)}\phi := \sum_{i=1}^n b_i \cdot D^{e_i}\phi + \frac{1}{2}\sum_{i,j=1}^n a_{ij}\cdot D^{e_i+e_j}\phi. \tag{27}$$

Assume that b and a are such that $b_i \cdot D^{e_i}\phi \in \mathcal{H}$ and $a_{ij}\cdot D^{e_i+e_j}\phi \in \mathcal{H}$ for $i,j = 1,\cdots,n$. This assumption holds, for example, when a and/or b are constants. With this assumption, Theorem 6 implies that $\mathrm{d}^{(2)}\phi \in \mathcal{H}$, and (26) can be written as $\big(\mathcal{A}^\star g\big)(x) = \langle g, \mathrm{d}^{(2)}\phi(x)\rangle$.

Therefore, (21) yields

$$\begin{aligned}\langle f, \partial C_0 g\rangle &= \mathbb{E}\Big[\big(\mathcal{A}^\star f\big)(X_0)g(X_0)\Big]\\ &= \mathbb{E}[\langle f, \mathrm{d}^{(2)}\phi(X_0)\rangle\langle g,\phi(X_0)\rangle]\\ &= \left\langle f\otimes g, \mathbb{E}\Big[\mathrm{d}^{(2)}\phi(X_0)\otimes\phi(X_0)\Big]\right\rangle\\ &= \left\langle f, \mathbb{E}\Big[\mathrm{d}^{(2)}\phi(X_0)\otimes\phi(X_0)\Big]g\right\rangle.\end{aligned} \tag{28}$$

Analogous to the covariance operator, ∂C_0 can be identified as the element in $\mathcal{H}_\otimes$ as

$$\partial C_0 := \mathbb{E}[\mathrm{d}^{(2)}\phi\left(X_0\right)\otimes\phi(X_0)], \tag{29}$$

The proof of the following result is in Appendix G.

Lemma 1. *Suppose Assumptions 1, 3, and 4 hold. Then* ∂C_0 *is a Hilbert-Schmidt operator.*

The PF generator is then $\mathcal{A} = \partial C_0 C_0^{-1}$. Similar to the discrete time case (c.f. (7)), its regularized variant is

$$\mathcal{A}_\varepsilon = \partial C_0\left(C_0 + \varepsilon\,\mathrm{id}\right)^{-1}. \tag{30}$$

7. Sparse Approximation of PF Generators

Writing $\mathcal{A}$ as an element in the tensor product space $\mathcal{H}_\otimes$ allows us to analyze the sample complexity of learning it from data along the lines of Tolstikhin et al. (2017, Proposition A.1) as follows. See Appendix H for proof.

Theorem 7. *Given* M *i.i.d. samples* $\{x_m\}_{m=1}^M$ *drawn according to* $\mathbb{P}_X$, *suppose Assumptions 1 and 4 hold. Define*

$$\widetilde{C}_0 := \frac{1}{M}\sum_{m=1}^M \varphi(x_m,x_m), \widetilde{\partial C}_0 := \frac{1}{M}\sum_{m=1}^M \mathrm{d}^{(2)}\phi\left(x_m\right)\otimes\phi\left(x_m\right),$$

and $\widetilde{\mathcal{A}}_\varepsilon := \widetilde{\partial C}_0\left(\widetilde{C}_0 + \varepsilon\, id\right)^{-1}$. *Then, for* $\delta \in (0,1)$, *the following holds with probability at least* $1-\delta$.

$$\left\|\partial C_0 - \widetilde{\partial C}_0\right\|_{HS} \leq \sqrt{\frac{B_\otimes}{M}}\Xi(\delta), \tag{31}$$

$$\left\|\widetilde{\mathcal{A}}_\varepsilon - \mathcal{A}_\varepsilon\right\| \leq \frac{B^{max}}{\sqrt{M}}\Xi(\delta/2)\mathcal{O}(\varepsilon^{-2}), \tag{32}$$

where $\sup_{x\in\mathbb{R}^n}\|\mathrm{d}^{(2)}\phi(x)\otimes\phi(x)\|_{H_\otimes}^2 \leq B_\otimes$ *with* $B_\otimes := B_\kappa^2\left(n^2+n^3+\frac{1}{4}n^4\right)\overline{S}$, *and* $B^{max} := \max\left\{\sqrt{B_\otimes}, B_\kappa\right\}$.

Next, we propose *sparse* learning of $\mathcal{A}_\varepsilon$ from i.i.d. data using the same notion of coherency introduced in (Mallat & Zhang, 1993; Tropp, 2004; Gilbert et al., 2003), but adapted to the setting involving the derivatives of functions in RKHS. Construct a γ-coherent dictionary by selecting a subset $\mathcal{D}_\gamma$ of $\mathcal{D}$ such that for each p, q with $x_p, x_q \in \mathcal{D}_\gamma$ and $p \neq q$,

$$\langle \mathrm{d}^{(2)}\phi(x_p), \mathrm{d}^{(2)}\phi(x_q)\rangle_{\mathcal{H}} \leq \sqrt{\gamma}\|\mathrm{d}^{(2)}\phi(x_p)\|_{\mathcal{H}}\|\mathrm{d}^{(2)}\phi(x_q)\|_{\mathcal{H}}, \tag{33}$$

and (11) with $(x_i^\star, x_j^\star) = (x_p, x_q)$. In addition, let $\mathcal{I}$ be the indices among $1, \ldots, M$ such that $x_i \in \mathcal{D}_\gamma$ for $i \in \mathcal{I}$. Then, the compressed estimator of ∂C_0 and C_0 are computed as

$$\widehat{\partial C}_0 = \sum_{i\in\mathcal{I}} z_i \mathrm{d}^{(2)}\phi(x_i)\otimes\phi(x_i), \widehat{C}_0 = \sum_{i\in\mathcal{I}} z_i' \varphi(x_i, x_i), \tag{34}$$

where z (and similarly, z') is obtained as $z = H^{-1}h$, with $H \in \mathbb{R}^{|\mathcal{D}_\gamma|\times|\mathcal{D}_\gamma|}$ has entries $H_{i,j} = \langle \mathrm{d}^{(2)}\phi(x_i)\otimes\phi(x_i), \mathrm{d}^{(2)}\phi(x_j)\otimes\phi(x_j)\rangle$ and $h \in \mathbb{R}^{|\mathcal{D}_\gamma|}$ is defined as $h_j = \frac{1}{M}\sum_{i=1}^M H_{i,j}$ for each i and j in $\mathcal{I}$. The sparse PF generator is then

$$\widehat{\mathcal{A}}_\varepsilon := \widehat{\partial C}_0 \left(\widehat{C}_0 + \varepsilon\,\mathrm{id}\right)^{-1}, \tag{35}$$

whose approximation error is given as follows. See Appendix I for a proof.

Theorem 8. *Under the assumptions of Theorem 7,* $\left\|\widehat{\mathcal{A}}_\varepsilon - \mathcal{A}_\varepsilon\right\| \leq B^{max}\psi_1(M, \gamma; \delta/2)\,\mathcal{O}(\varepsilon^{-2})$ *holds with probability at least* $1-\delta$ *for any* $\delta \in (0,1)$*, where* ψ_1 *is defined in* (13)*.*

Similar to the discrete time case, one can learn the generator from trajectory data. We consider a generalization of the notion of β-mixing in discrete-time to continuous-time as follows. Define the β-coefficients $\beta(s)$ exactly as (14), and call the process $\{X_t\}_{t\in\mathbb{T}}$ β-mixing, if $\beta(s) \to 0$ as $s \to \infty$[3]. A direct application of results from Section 4 gives the following result.

Theorem 9. *Let* $\{X_t\}_{t\in\mathbb{T}}$ *be a* β*-mixing diffusion process defined on a filtered probability space* $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\in\mathbb{T}}, \mathbb{P})$ *where* X_t *is* $\mathcal{F}_t$*-adapted and takes values in* $\mathbb{X} = \mathbb{R}^n$*. Suppose it has a unique stationary distribution* Π *and satisfies Assumption 2. Consider the evolution of the state, starting from* X_0 *that is sampled according to an initial distribution* $\mathbb{P}_{X_0}$*. Let* $X^s(m) = \left(X_{(1)}, \ldots, X_{(m)}\right)$ *be an* m*-length sequence, sampled* s *time apart. Under Assumptions 1, 2 and 4,* $\left\|\widehat{\mathcal{A}}_\varepsilon - \mathcal{A}_\varepsilon\right\| \leq B^{max}\psi_2(m, s, \gamma; \delta/2)\mathcal{O}(\varepsilon^{-2})$ *holds with probability at least* $1-\delta$ *for* $m\beta(s) < \delta < 1$*.*

[3]When restricted to stationary Markov processes, an alternate but equivalent definition is given by Aït-Sahalia et al. (2010, Definition 8)

8. Learning SDE Coefficients from Data

Our data-driven approximation of the regularized PF generator $\mathcal{A}_\varepsilon$ in Section 7 is premised on the knowledge of the drift and diffusion coefficients, b and $a = \sigma\sigma^\top$, respectively, of the SDE in (24). We now study the same, when b and a are not known. Notice that

$$\begin{aligned} b(x) &= \lim_{\tau\to 0}\frac{1}{\tau}\mathbb{E}\left[(X_\tau - X_0)|X_0 = x\right], \\ a(x) &= \lim_{\tau\to 0}\frac{1}{\tau}\mathbb{E}\left[(X_\tau - X_0)(X_\tau - X_0)^\top|X_0 = x\right]. \end{aligned} \tag{36}$$

We approximate components of b and a using their finite difference approximations, $b^\tau(x)$ and $a^\tau(x)$, respectively. To motivate the key idea, consider a scalar-valued process, i.e., $X_t \in \mathbb{R}$. With a slight abuse of notation, assume that the identity function $\mathrm{id} : \mathbb{R} \to \mathbb{R}$ is an element of $\mathcal{H}$. For any $t \geq 0$, we have $\mathbb{E}[X_t|X_0 = x] = \langle \mathrm{id}, \mathbb{E}[\kappa(X_t, \cdot)|X_0 = x]\rangle$, where $\mathbb{E}[\kappa(X_t,\cdot)|X_0 = x]$ is the embedding of the distribution of X_t in $\mathcal{H}$, given $X_0 = x$. At $t = 0$, this equates to $\kappa(x,\cdot)$, and at $t = \tau$, to $\mathcal{U}_\tau\kappa(x,\cdot)$. Taken together, these observations yield

$$\tau b^\tau(x) = \langle \mathrm{id}, \mathcal{U}_\tau\kappa(x,\cdot) - \kappa(x,\cdot)\rangle = \langle \mathcal{U}_\tau^\star\mathrm{id} - \mathrm{id}, \kappa(x,\cdot)\rangle,$$

implying $b^\tau = \frac{1}{\tau}(\mathcal{U}_\tau^\star\mathrm{id} - \mathrm{id})$. Likewise, $\mathbb{E}\left[\kappa(X_t,\cdot)\,\kappa(X_{t'},\cdot)\,|X_0 = x\right]$ is the embedding of the conditional joint distribution of $(X_t, X_{t'})$, given $X_0 = x$. Proceeding similarly as before, this equals $\langle \mathrm{id}\otimes\mathrm{id}, \mathbb{E}[\kappa(X_t,\cdot)\kappa(X_{t'},\cdot)|X_0 = x]\rangle$, where the second term in the inner product can be written as $\mathcal{U}_{tt'}\kappa(x,\cdot)$. Here, $\mathcal{U}_{tt'}$ denotes the conditional joint mean embedding operator that generalizes the CME $\mathcal{U}_t$ to the tensor product Hilbert space. We defer its formal definition to Appendix J and remark that under a condition similar to that in Assumption 3, $\mathcal{U}_{tt'}$ can be defined using joint covariance operators $C_{tt'0}$ and covariance operator C_0, also defined formally by (98) in Appendix J. Using this notation,

$$\begin{aligned} \tau a^\tau(x) =& \mathbb{E}\left[X_\tau^2 - 2X_0X_\tau + X_0^2|X_0 = x\right] \\ =& \big\langle \mathcal{U}_{\tau\tau}^\star(\mathrm{id}\otimes\mathrm{id}) - 2\mathcal{U}_{\tau 0}^\star(\mathrm{id}\otimes\mathrm{id}) \\ &+ (\mathrm{id}\otimes\mathrm{id})\kappa(x,\cdot), \kappa(x,\cdot)\big\rangle_{\mathcal{H}}, \end{aligned} \tag{37}$$

implying $\tau a^\tau = \mathcal{U}_{\tau\tau}^\star(\mathrm{id}\otimes\mathrm{id}) - 2\mathcal{U}_{\tau 0}^\star(\mathrm{id}\otimes\mathrm{id}) + (\mathrm{id}\otimes\mathrm{id})\kappa(x,\cdot)$. The conditional joint mean embedding operators can be estimated from $\{x_m(0), x_m(\tau)\}_{m=1}^N$ of N, drawn i.i.d. according to $\mathbb{P}(X_0, X_\tau)$, where $x_m(\tau)$ is the next snapshot of $x_m(0)$ with time lag τ. The details of the estimation of (regularized) joint covariance and conditional joint mean embedding operators from data are relegated to Appendix J. Call the estimates of b^τ and a^τ with estimated $\mathcal{U}$'s as $\widehat{b^\tau}$ and $\widehat{a^\tau}$, respectively.

For an n-dimensional diffusion process, we apply the same technique to each scalar component of b and a. Let $\boldsymbol{e}_i$:

$X \mapsto X^{(i)}$ evaluate the i-th coordinate of $X \in \mathbb{R}^n$ and assume that $\boldsymbol{e}_i \in \mathcal{H}$ for $i = 1, \cdots, n$. Then, we have

$$\begin{aligned} b_i^\tau =& \frac{1}{\tau}\left(\mathcal{U}_\tau^\star \boldsymbol{e}_i - \boldsymbol{e}_i\right), \\ a_{ij}^\tau =& \frac{1}{\tau}\left(\mathcal{U}_{\tau\tau}^\star(\boldsymbol{e}_i \otimes \boldsymbol{e}_j) - \mathcal{U}_{\tau 0}^\star(\boldsymbol{e}_i \otimes \boldsymbol{e}_j)\right. \\ & \left. - \mathcal{U}_{\tau 0}^\star(\boldsymbol{e}_j \otimes \boldsymbol{e}_i) + (\boldsymbol{e}_i \otimes \boldsymbol{e}_j)\kappa(x, \cdot)\right). \end{aligned} \tag{38}$$

Using empirical estimates $\widehat{b_i^\tau}$, $\widehat{a_{ij}^\tau}$ of the drift and the diffusion coefficients, an empirical estimate of ∂C_0 can be constructed from data. In turn, the estimated ∂C_0 yield empirical generators. Such an entirely data-driven algorithm is summarized in Appendix L.

The development in this section requires $\mathrm{id} \in \mathcal{H}$. An infinite-dimensional RKHS with bounded kernel cannot include such a map (Muandet et al., 2016, Section 2.3). Nevertheless, we view the expressions to evaluate $\widehat{a^\tau}$ and $\widehat{b^\tau}$ as ways to obtain an approximation, where the actions of id can be directly utilized without explicitly representing this map within the RKHS. See Appendix K for details.

9. Numerical Experiments

In this section, we report results from applications of our algorithms to approximations of eigenfunctions of $\mathcal{A}$ and $\mathcal{A}^\star$. Eigenfunction construction using Gram matrices is described in Appendix M. When the SDE coefficients are not known, one can first estimate them following Section 8 using a data set of snapshot pairs $\mathcal{D}_1$. The generator is then constructed as in Section 7, utilizing a second dataset $\mathcal{D}_2 = \{x_m\}_{m=1}^M$. A more detailed description is provided in Appendix L.

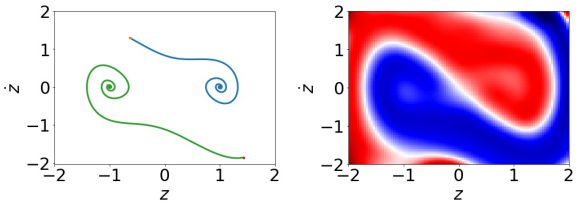
9.1. The Duffing Oscillator

Consider the unforced Duffing oscillator, described by

$$\ddot{z} = -\delta\dot{z} - z\left(\beta + \alpha z^2\right),$$

with $\delta = 0.5$, $\beta = -1$, and $\alpha = 1$, where $z \in \mathbb{R}$ and $\dot{z} \in \mathbb{R}$ are the scalar position and velocity, respectively. Let $x = (z, \dot{z})$. As Figure 1a reveals, this system exhibits two regions of attraction, corresponding to equilibrium points $x = (-1, 0)$ and $x = (1, 0)$.

To approximate $\mathcal{A}^\star$, we sample 100 trajectories to form $\mathcal{D}_1$ by first generating 100 uniformly distributed initial points from $[-2, 2] \times [-2, 2]$, then propagating them through the dynamics by evolving 1000 steps with sampling interval $\tau = 0.01$ s. We then create a sub-sample with $m = 200$, and $s = 5$ along each trajectory. We utilize a Gaussian kernel $\kappa(x_1, x_2) = \exp(-\|x_1 - x_2\|_2^2/(2 \times 0.65^2))$, whose partial derivatives are included in Appendix N. Setting $\gamma_1 = 0.99^2$, we get $|\mathcal{D}_{\gamma_1}| = 1477$. Likewise, $\mathcal{D}_2$ is constructed from 50 uniformly distributed initial points with 100 evaluations along each and then sub-sampled with $m = 20, s = 5$ for each trajectory. Using $\gamma_2 = 0.9995^2$, we obtain $|\mathcal{D}_{\gamma_2}| = 876$. Figure 1b portrays heat-maps of the leading eigenfunctions of $\mathcal{A}^\star$ with learned coefficients. We refer interested readers to (Hou et al., 2021) for a discussion of the practical benefits of sparsification and the role of γ.



(a) (b)

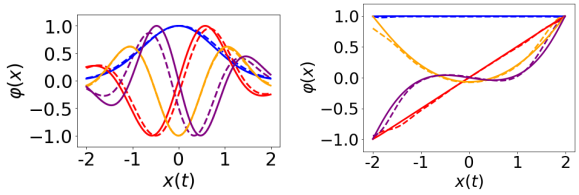
Figure 1: (a) Two trajectories of the Duffing oscillator that converge to two different equilibrium points. (b) Leading eigenfunction of $\widehat{\mathcal{A}_\varepsilon^\star}$ with eigenvalue 0.

9.2. One-Dimensional Ornstein-Uhlenbeck Process

Consider a one-dimensional Ornstein-Uhlenbeck process defined by the SDE

$$\mathrm{d}X_t = -\alpha D X_t \mathrm{d}t + \sqrt{2D}\mathrm{d}B_t,$$

with $\alpha = 4$, $D = \frac{1}{4}$. With explicit knowledge of the drift and diffusion coefficients, we first create a dataset $\mathcal{D}$ from 10 trajectories with 5000 evaluations each using sampling interval $\tau = 0.1$s. The evolution was accomplished through 100 steps of the Euler-Maruyama method (Higham, 2001) with a time-step of 10^{-3}. We then form a sub-sample by choosing $m = 50$ and $s = 100$ for each trajectory. Using Gaussian kernel $\kappa(x_1, x_2) = \exp(-\|x_1 - x_2\|_2^2/(2 \times 0.6^2))$, and setting $\gamma = 0.999^2$, we obtain $\mathcal{D}_\gamma$ which contains 48 samples. Figure 2 shows the leading eigenfunctions of sparse estimates of $\mathcal{A}, \mathcal{A}^\star$



(a) (b)

Figure 2: Leading eigenfunctions φ of sparse empirical estimates of $\mathcal{A}$, $\mathcal{A}^\star$ with (solid line)/ without (dashed line) explicit knowledge of drift and diffusion coefficients corresponding to eigenvalues $\lambda = 0$(—), -1(—), -2(—), -3(—).

When the drift and diffusion coefficients are unknown, we first sample 200 initial points uniformly distributed over

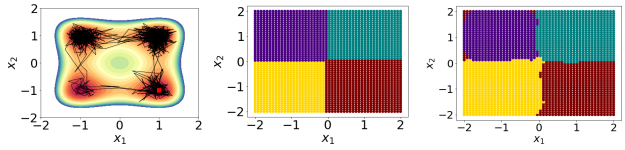
$[-2, 2]$, then collect $10,000$ points along each trajectory with sampling interval 0.1s. We next create a sub-sample of size $20,000$ with $m = 100, s = 100$ for each trajectory. $\mathcal{D}_2$ is chosen to be the same dataset as $\mathcal{D}$ described in the previous paragraph. A Gaussian kernel $\kappa(x_1, x_2) = \exp(-\|x_1 - x_2\|_2^2/(2 \times 0.6^2))$ is used. With $\gamma_1 = 0.999, \gamma_2 = 0.9999^2$, we obtain $|\mathcal{D}_{\gamma_1}| = 1325$ and $|\mathcal{D}_{\gamma_2}| = 106$. The first 4 eigenfunctions of $\mathcal{A}$ and $\mathcal{A}^\star$ are shown in Figure 2.

9.3. A Two-Dimensional Quadruple-Well

Consider the SDE for the quadruple-well dynamics,

$$\mathrm{d}X_t = -\nabla V(X_t)\mathrm{d}t + \sqrt{2\beta^{-1}}\mathrm{d}B_t,$$

with $x = [x_1, x_2], V(x) = (x_1^2-1)^2+(x_2^2-1)^2$ and $\beta = 4$. As illustrated in Figure 3a, a trajectory will stay within one of the four potential wells, while rare transitions happen as "jumps" between four metastable sets.



(a) (b) (c)

Figure 3: (a) Potential landscape and one trajectory of the quadruple-well dynamics; (b),(c) four metastable set obtained from leading eigenfunctions of sparse $\mathcal{A}^\star$ with (middle)/ without (right) explicit knowledge of drift and diffusion terms.

We first approximate $\mathcal{A}^\star$ and its leading eigenfunctions with explicit knowledge of drift and diffusion. $\mathcal{D}$ consists of samples on $[-2, 2] \times [-2, 2]$ collected from 600 trajectories with 20 evolutions along each with sampling interval $\tau = 0.1s$. A sub-sample is then constructed with $m = 2, s = 10$ for each trajectory. We use $\kappa(x_1, x_2) = \exp(-\|x_1 - x_2\|_2^2/(2 \times 0.2^2))$. With $\gamma = 0.81$, we get $|\mathcal{D}_\gamma| = 604$. Since the spectrum encodes state space connectivity information, we then apply k-means clustering techniques to dominant eigenfunctions (Froyland et al., 2019) to locate metastable sets which are shown in Figure 3b.

Next, we consider the case when neither drift nor diffusion coefficients are known. We create $\mathcal{D}_1$ by first collecting from 1200 trajectories with $5,000$ evaluations each with sampling interval $\tau = 0.1$s. We then construct a sub-sample of size $120,000$ with $m = 100, s = 50$ for each trajectory. We utilize $\mathcal{D}$ described in the previous paragraph as $\mathcal{D}_2$. Using the same kernel functions with $\gamma_1 = 0.8, \gamma_2 = 0.81$, we obtain $\mathcal{D}_{\gamma_1}, \mathcal{D}_{\gamma_2}$ with $|\mathcal{D}_{\gamma_1}| = 7484$ and $|\mathcal{D}_{\gamma_2}| = 594$. The resulting four metastable sets are shown in Figure 3c.

Remark 1. *In our numerical experiments, Assumption 3 on the closedness of RKHS under the action of the system dynamics is challenging to verify. It possibly does not hold with Gaussian kernels (see discussions in Klebanov et al. (2020); Park & Muandet (2020)). Yet, the results demonstrate that, as a computational method, sparse learning of CME performs well, even when the assumptions made for the theoretical analyses are violated. On a related note, notice that the Ornstein-Uhlenbeck process has a stationary distribution, and its stationary variant is β-mixing, according to Meyn & Tweedie (1993), Jongbloed et al. (2005, Section 3). By contrast, the Duffing oscillator has two stable equilibrium points with two different regions of attraction, indicating that Assumption 2 does not hold. Yet, sparse estimates of transfer operators/generators provide efficient computational techniques to analyze dynamical system properties.*

10. Conclusions

In this paper, we have provided sample complexity bounds for approximations of transfer operators from data that is sparsified and collected from trajectories. We have rigorously defined and characterized the generators of transfer operators for continuous-time Markov processes using partial derivatives of kernel functions and covariance-type operators. Then, we have provided sample complexity bounds for approximating these generators. Numerical experiments confirmed the effectiveness of our approach.

An interesting direction for future work is the study of generators of transfer operators for *controlled* diffusion processes. Particularly, for continuous-time Markov decision processes, we plan to approximate PF generators parameterized by control policies to ultimately approximate solutions of Hamilton–Jacobi–Bellman equations in an RKHS. We also plan to extend our approach to tackle the online streaming setting where samples are collected sequentially, possibly in a decentralized fashion by multiple agents.

Acknowledgements

This work was supported by the NSF-EPCN-2031570 grant, NSF-CPS-2038775 grant and C3.ai Digital Transformation Institute. The authors thank the anonymous reviewers for their insightful comments.

References

Agarwal, A. and Duchi, J. C. The generalization ability of online algorithms for dependent data. *IEEE Transactions on Information Theory*, 59(1):573–587, 2012.

Aït-Sahalia, Y., Hansen, L. P., and Scheinkman, J. A. Operator methods for continuous-time markov processes. *Handbook of financial econometrics: tools and techniques*, pp. 1–66, 2010.

Anthony, M., Bartlett, P. L., Bartlett, P. L., et al. *Neural network learning: Theoretical foundations*, volume 9. cambridge University Press, 1999.

Aubin, J.-P. *Applied functional analysis*. John Wiley & Sons, 2011.

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

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

Fischer, S. and Steinwart, I. Sobolev norm learning rates for regularized least-squares algorithms. *The Journal of Machine Learning Research*, 21(1):8464–8501, 2020.

Froyland, G., Rock, C. P., and Sakellariou, K. Sparse eigenbasis approximation: Multiple feature extraction across spatiotemporal scales with application to coherent set identification. *Communications in Nonlinear Science and Numerical Simulation*, 77:81–107, 2019.

Fukumizu, K., Bach, F. R., and Jordan, M. I. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *Journal of Machine Learning Research*, 5 (Jan):73–99, 2004.

Fukumizu, K., Bach, F. R., and Gretton, A. Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 8(2), 2007.

Gilbert, A. C., Muthukrishnan, S., and Strauss, M. J. Approximation of functions over redundant dictionaries using coherence. In *SODA*, pp. 243–252. Citeseer, 2003.

Gretton, A. Notes on mean embeddings and covariance operators, 2015.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Higham, D. J. An algorithmic introduction to numerical simulation of stochastic differential equations. *SIAM review*, 43(3):525–546, 2001.

Hou, B., Bose, S., and Vaidya, U. Sparse learning of kernel transfer operators. In *2021 55th Asilomar Conference on Signals, Systems, and Computers*, pp. 130–134. IEEE, 2021.

Jongbloed, G., Van Der Meulen, F. H., and Van Der Vaart, A. W. Nonparametric inference for lévy-driven ornstein-uhlenbeck processes. *Bernoulli*, 11(5):759–791, 2005.

Kivinen, J., Smola, A. J., and Williamson, R. C. Online learning with kernels. *IEEE transactions on signal processing*, 52(8):2165–2176, 2004.

Klebanov, I., Schuster, I., and Sullivan, T. J. A rigorous theory of conditional mean embeddings. *SIAM Journal on Mathematics of Data Science*, 2(3):583–606, 2020.

Klus, S., Nüske, F., and Hamzi, B. Kernel-based approximation of the koopman generator and schrödinger operator. *Entropy*, 22(7):722, 2020a.

Klus, S., Nüske, F., Peitz, S., Niemann, J.-H., Clementi, C., and Schütte, C. Data-driven approximation of the koopman generator: Model reduction, system identification, and control. *Physica D: Nonlinear Phenomena*, 406: 132416, 2020b.

Klus, S., Schuster, I., and Muandet, K. Eigendecompositions of transfer operators in reproducing kernel hilbert spaces. *Journal of Nonlinear Science*, 30(1):283–315, 2020c.

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

Kostic, V., Novelli, P., Maurer, A., Ciliberto, C., Rosasco, L., and Pontil, M. Learning dynamical systems via koopman operator regression in reproducing kernel hilbert spaces. *arXiv preprint arXiv:2205.14027*, 2022.

Kuznetsov, V. and Mohri, M. Generalization bounds for time series prediction with non-stationary processes. In *International Conference on Algorithmic Learning Theory*, pp. 260–274. Springer, 2014.

Lever, G., Shawe-Taylor, J., Stafford, R., and Szepesvári, C. Compressed conditional mean embeddings for model-based reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

Li, Z., Meunier, D., Mollenhauer, M., and Gretton, A. Optimal rates for regularized conditional mean embedding learning. *arXiv preprint arXiv:2208.01711*, 2022.

Lopez-Paz, D., Muandet, K., Schölkopf, B., and Tolstikhin, I. Towards a learning theory of cause-effect inference. In *International Conference on Machine Learning*, pp. 1452–1461. PMLR, 2015.

Mallat, S. G. and Zhang, Z. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

Mauroy, A. and Goncalves, J. Koopman-based lifting techniques for nonlinear systems identification. *IEEE Transactions on Automatic Control*, 65(6):2550–2565, 2019.

Meyn, S. P. and Tweedie, R. L. Stability of markovian processes iii: Foster–lyapunov criteria for continuous-time processes. *Advances in Applied Probability*, 25(3): 518–548, 1993.

Mohri, M. and Rostamizadeh, A. Rademacher complexity bounds for non-iid processes. *Advances in Neural Information Processing Systems*, 21, 2008.

Mollenhauer, M., Klus, S., Schütte, C., and Koltai, P. Kernel autocovariance operators of stationary processes: Estimation and convergence. *arXiv preprint arXiv:2004.00891*, 2020.

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

Nandanoori, S. P., Sinha, S., and Yeung, E. Data-driven operator theoretic methods for global phase space learning. In *2020 American Control Conference (ACC)*, pp. 4551–4557. IEEE, 2020.

Nüske, F., Peitz, S., Philipp, F., Schaller, M., and Worthmann, K. Finite-data error bounds for koopman-based prediction and control. *Journal of Nonlinear Science*, 33 (1):14, 2023.

Park, J. and Muandet, K. A measure-theoretic approach to kernel conditional mean embeddings. *Advances in Neural InformationProcessing Systems*, 33:21247–21259, 2020.

Richard, C., Bermudez, J. C. M., and Honeine, P. Online prediction of time series data with kernels. *IEEE Transactions on Signal Processing*, 57(3):1058–1067, 2008.

Rosenfeld, J. A., Russo, B., Kamalapurkar, R., and Johnson, T. T. The occupation kernel method for nonlinear system identification. *arXiv preprint arXiv:1909.11792*, 2019.

Schilling, R. L. Brownian motion. In *Brownian Motion*. de Gruyter, 2021.

Smola, A., Gretton, A., Song, L., and Schölkopf, B. A hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory*, pp. 13–31. Springer, 2007.

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

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

Steinwart, I. and Scovel, C. Mercer's theorem on general domains: On the interaction between measures, kernels, and rkhss. *Constructive Approximation*, 35:363–417, 2012.

Tolstikhin, I., Sriperumbudur, B. K., and Muandet, K. Minimax estimation of kernel mean embeddings. *The Journal of Machine Learning Research*, 18(1):3002–3048, 2017.

Tropp, J. A. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information theory*, 50(10):2231–2242, 2004.

Vidyasagar, M. *Learning and generalisation: with applications to neural networks*. Springer Science & Business Media, 2013.

Vrabie, I. I. *Co-semigroups and applications*. Elsevier, 2003.

Williams, M. O., Rowley, C. W., and Kevrekidis, I. G. A kernel-based approach to data-driven koopman spectral analysis. *Journal of Computational Dynamics*, 2014.

Williams, M. O., Kevrekidis, I. G., and Rowley, C. W. A data–driven approximation of the koopman operator: Extending dynamic mode decomposition. *Journal of Nonlinear Science*, 25(6):1307–1346, 2015.

Wu, M., Schölkopf, B., Bakır, G., and Cristianini, N. A direct method for building sparse kernel learning algorithms. *Journal of Machine Learning Research*, 7(4), 2006.

Yeung, E., Kundu, S., and Hodas, N. Learning deep neural network representations for koopman operators of nonlinear dynamical systems. In *2019 American Control Conference (ACC)*, pp. 4832–4839. IEEE, 2019.

Yu, B. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, pp. 94–116, 1994.

Zhou, D.-X. Derivative reproducing properties for kernel methods in learning theory. *Journal of Computational and Applied Mathematics*, 220(1-2):456–463, 2008.

A. Proof of Theorem 1

Triangle inequality and elementary algebra gives

$$\begin{aligned}\left\|\widehat{\mathcal{P}}_\varepsilon - \mathcal{P}_\varepsilon\right\| &\leq \left\|\widehat{C}_{X^+X}\left(\widehat{C}_{XX}+\varepsilon\,\mathrm{id}\right)^{-1} - C_{X^+X}\left(\widehat{C}_{XX}+\varepsilon\,\mathrm{id}\right)^{-1}\right\| \\ &\quad + \left\|C_{X^+X}\left(\widehat{C}_{XX}+\varepsilon\,\mathrm{id}\right)^{-1} - C_{X^+X}\left(C_{XX}+\varepsilon\,\mathrm{id}\right)^{-1}\right\| \\ &= \left\|\left(\widehat{C}_{X^+X}-C_{X^+X}\right)\left(\widehat{C}_{XX}+\varepsilon\,\mathrm{id}\right)^{-1}\right\| \\ &\quad + \left\|C_{X^+X}\left[\left(\widehat{C}_{XX}+\varepsilon\,\mathrm{id}\right)^{-1} - \left(C_{XX}+\varepsilon\,\mathrm{id}\right)^{-1}\right]\right\|.\end{aligned} \tag{39}$$

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

$$\begin{aligned}Z_1 &\overset{(a)}{\leq} \left\|\widehat{C}_{X^+X}-C_{X^+X}\right\|\left\|\left(\widehat{C}_{XX}+\varepsilon\,\mathrm{id}\right)^{-1}\right\| \\ &\overset{(b)}{\leq} \left\|\widehat{C}_{X^+X}-C_{X^+X}\right\|_{\text{HS}}\left\|\left(\widehat{C}_{XX}+\varepsilon\,\mathrm{id}\right)^{-1}\right\| \\ &\overset{(c)}{\leq} \frac{1}{\varepsilon}\left\|\widehat{C}_{X^+X}-C_{X^+X}\right\|_{\text{HS}}.\end{aligned} \tag{40}$$

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 positive semi-definite [4] and self-adjoint, where the latter property implies that the operator norm coincides with their spectral radius, thus we have

$$\left\|\left(\widehat{C}_{XX}+\varepsilon\,\mathrm{id}\right)^{-1}\right\| \leq \frac{1}{\varepsilon}. \tag{41}$$

Proceeding similarly, we bound Z_2 as

$$\begin{aligned}Z_2 &\leq \|C_{X^+X}\|\left\|\left(\widehat{C}_{XX}+\varepsilon\,\mathrm{id}\right)^{-1} - \left(C_{XX}+\varepsilon\,\mathrm{id}\right)^{-1}\right\| \\ &= \|C_{X^+X}\| \times \left\|\left(\widehat{C}_{XX}+\varepsilon\,\mathrm{id}\right)^{-1}\left(C_{XX}-\widehat{C}_{XX}\right)\left(C_{XX}+\varepsilon\,\mathrm{id}\right)^{-1}\right\| \\ &\leq \frac{1}{\varepsilon^2}\|C_{X^+X}\|\left\|C_{XX}-\widehat{C}_{XX}\right\| \\ &\leq \frac{1}{\varepsilon^2}\|C_{X^+X}\|\left\|C_{XX}-\widehat{C}_{XX}\right\|_{\text{HS}}.\end{aligned} \tag{42}$$

Here, the second last line follows from using the relation (41) and its counterpart with $\widehat{C}_{XX}$ replaced by C_{XX}. Utilizing (40) and (42) in (39), we get

$$\left\|\widehat{\mathcal{P}}_\varepsilon - \mathcal{P}_\varepsilon\right\| \leq \frac{1}{\varepsilon}\left\|\widehat{C}_{X^+X}-C_{X^+X}\right\|_{\text{HS}} + \frac{1}{\varepsilon^2}\|C_{X^+X}\|\left\|C_{XX}-\widehat{C}_{XX}\right\|_{\text{HS}}. \tag{43}$$

Define the non-sparse empirical estimates of the covariance and cross-covariance operators as

$$\widetilde{C}_{XX} := \frac{1}{M}\sum_{i=1}^{M}\varphi(x_i,x_i), \quad \widetilde{C}_{X^+X} := \frac{1}{M}\sum_{i=1}^{M}\varphi(x_i^+,x_i). \tag{44}$$

[4] See the remark in Theorem 1

Using triangle inequality, we get

$$\left\|\widehat{C}_{X^+X} - C_{X^+X}\right\|_{\text{HS}} \leq \left\|\widetilde{C}_{X^+X} - C_{X^+X}\right\|_{\text{HS}} + \left\|\widehat{C}_{X^+X} - \widetilde{C}_{X^+X}\right\|_{\text{HS}}. \tag{45}$$

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 (45).

First, note that the HS norm of a (cross) covariance operator from $\mathcal{H}$ to $\mathcal{H}$ is equal to the $\mathcal{H}_\otimes$-norm of the operator viewed as an element in tensor product Hilbert space $\mathcal{H}_\otimes$ Fukumizu et al. (2007, Lemma 4):

$$\|C_{X^+X}\|_{\text{HS}}^2 = \left\|\mathbb{E}_{X^+X}\left[\kappa(\cdot, X^+)\kappa(\cdot, X)\right]\right\|_{\mathcal{H}_\otimes}^2 \tag{46}$$

Moreover, recall that C_{X^+X} and C_{XX} are the embeddings of $\mathbb{P}(X, X^+)$ and its marginal $\mathbb{P}_X$ in $\mathcal{H}_\otimes$. Under Assumption 1, we have

$$\sup_x \kappa^2(x, x) \leq \left(\sup_x \kappa(x, x)\right)\left(\sup_x \kappa(x, x)\right) \leq B_\kappa^2 < \infty.$$

Together with Muandet et al. (2016, Theorem 3.4), we bound the first term on the right-hand-side of (45) as

$$\left\|\widetilde{C}_{X^+X} - C_{X^+X}\right\|_{\text{HS}} \leq B_\kappa/\sqrt{M}\left(1 + \sqrt{2\log(1/\delta)}\right) \tag{47}$$

with probability at least $1 - \delta$. The same bound applies to $\widetilde{C}_{XX} - C_{XX}$. For bounding the second term on the right-hand side of (45), we introduce additional notation. Recalling that $\mathcal{I}$ is the set of indices among $\mathcal{D}$ that are present in $\mathcal{D}_\gamma$, let $\Pi_{\mathcal{D}_\gamma}$ be the (linear) projection operator on the closed subspace $\{\varphi(x_i^+, x_i) : i \in \mathcal{I}\}$ of $\mathcal{H}_\otimes$. Then, we have

$$\begin{aligned}
\left\|\widehat{C}_{X^+X} - \widetilde{C}_{X^+X}\right\|_{\text{HS}} &= \left\|\frac{1}{M}\sum_{i=1}^{M}\left(\text{id} - \Pi_{\mathcal{D}_\gamma}\right)\varphi\left(x_i^+, x_i\right)\right\|_{\mathcal{H}_\otimes} \\
&\leq \sum_{i=1}^{M}\frac{1}{M}\left\|\varphi\left(x_i^+, x_i\right) - \Pi_{\mathcal{D}_\gamma}\varphi(x_i^+, x_i)\right\|_{\mathcal{H}_\otimes} \\
&= \frac{1}{M}\sum_{i\notin\mathcal{I}}\left\|\varphi\left(x_i^+, x_i\right) - \Pi_{\mathcal{D}_\gamma}\varphi\left(x_i^+, x_i\right)\right\|_{\mathcal{H}_\otimes}.
\end{aligned} \tag{48}$$

Pythagoras' theorem gives

$$\left\|\varphi\left(x_i^+, x_i\right) - \Pi_{\mathcal{D}_\gamma}\varphi\left(x_i^+, x_i\right)\right\|_{\mathcal{H}_\otimes}^2 = \underbrace{\left\|\varphi\left(x_i^+, x_i\right)\right\|_{\mathcal{H}_\otimes}^2}_{T_1} - \underbrace{\left\|\Pi_{\mathcal{D}_\gamma}\varphi\left(x_i^+, x_i\right)\right\|_{\mathcal{H}_\otimes}^2}_{T_2}, \tag{49}$$

where $T_1 \leq B_\kappa^2$. From the coherence condition (11), we have

$$\frac{\left|\kappa_\otimes\Big((x_i^+, x_i), (x_j^+, x_j)\Big)\right|}{\sqrt{\kappa_\otimes\Big((x_i^+, x_i), (x_i^+, x_i)\Big)\kappa_\otimes\Big((x_j^+, x_j), (x_j^+, x_j)\Big)}} \leq \frac{\left|\kappa(x_i^+, x_j^+)\right|}{\sqrt{\kappa(x_i^+, x_i^+)\kappa(x_j^+, x_j^+)}}\frac{|\kappa(x_i, x_j)|}{\sqrt{\kappa(x_i, x_i)\kappa(x_j, x_j)}} \leq \gamma, \tag{50}$$

for each pair (i,j) such that (x_i, x_i^+) and (x_j, x_j^+) are in $\mathcal{D}_\gamma$. Thus we can bound T_2 from below as

$$\begin{aligned}\left\|\Pi_{\mathcal{D}_\gamma}\varphi(x_i^+,x_i)\right\| &= \max_{\alpha}\left\langle \frac{\sum_{j\in\mathcal{I}}\alpha_j\varphi(x_j^+,x_j)}{\left\|\sum_{j\in\mathcal{I}}\alpha_j\varphi(x_j^+,x_j)\right\|_{H_\otimes}}, \varphi(x_i^+,x_i)\right\rangle_{\mathcal{H}_\otimes} \\ &= \max_{\alpha}\frac{\sum_{j\in\mathcal{I}}\alpha_j\kappa_\otimes\Big((x_i^+,x_i),(x_j^+,x_j)\Big)}{\left\|\sum_{j\in\mathcal{I}}\alpha_j\varphi(x_j^+,x_j)\right\|_{H_\otimes}} \\ &\overset{(a)}{\geq} \max_{q\in\mathcal{I}}\frac{\left|\kappa_\otimes\Big((x_i^+,x_i),(x_q^+,x_q)\Big)\right|}{\sqrt{\kappa_\otimes\Big((x_q^+,x_q),(x_q^+,x_q)\Big)}} \\ &\overset{(b)}{\geq} \gamma\sqrt{\kappa_\otimes\Big((x_i^+,x_i),(x_i^+,x_i)\Big)},\end{aligned} \tag{51}$$

where (a) results from a specific choice of coefficients. Specifically, it is obtained with $\alpha_j = 0$ for each $j \in \mathcal{I}$, expect for a single index q with $\alpha_q = \pm 1$, depending on the sign of $\kappa_\otimes\Big((x_i^+,x_i),(x_q^+,x_q)\Big)$. From the violation of the coherence condition (11), we obtain (b). Combining the bounds on T_1 and T_2 in (49), we get

$$\begin{aligned}\left\|\varphi\left(x_i^+,x_i\right) - \Pi_{\mathcal{D}_\gamma}\varphi\left(x_i^+,x_i\right)\right\|^2 &\leq \kappa_\otimes\left((x_i^+,x_i),(x_i^+,x_i)\right)(1-\gamma^2) \\ &\leq B_\kappa^2(1-\gamma^2).\end{aligned} \tag{52}$$

Utilizing this bound in (48), we get

$$\left\|\widehat{C}_{X^+X} - \widetilde{C}_{X^+X}\right\|_{\text{HS}} \leq \left(1 - \frac{|\mathcal{D}_\gamma|}{M}\right) B_\kappa\sqrt{1-\gamma^2}. \tag{53}$$

Combining (47) and (53) in (45), we obtain

$$\left\|\widehat{C}_{X^+X} - C_{X^+X}\right\|_{\text{HS}} \leq B_\kappa\psi_1(M,\gamma;\delta) \tag{54}$$

with probability $\geq 1-\delta$. The same bound applies to the sparse approximation $\widehat{C}_{XX}$ of C_{XX}. We then split the failure probability δ evenly between the cross covariance and covariance estimations, and use union bound to obtain the required result as

$$\left\|\widehat{\mathcal{P}}_\varepsilon - \mathcal{P}_\varepsilon\right\| \leq B_\kappa\psi_1(M,\gamma;\delta/2)\left(\frac{1}{\varepsilon} + \frac{\|C_{X^+X}\|}{\varepsilon^2}\right), \tag{55}$$

B. Proof of Theorem 2

Consider two points (x^+, x) and (x'^+, x') in $\mathcal{D}_\gamma$. Then, we have

$$\begin{aligned}\|\varphi(x^+,x) - \varphi(x'^+,x')\|_{\mathcal{H}_\otimes}^2 &= \kappa_\otimes\Big((x^+,x),(x^+,x)\Big) + \kappa_\otimes\Big((x'^+,x'),(x'^+,x')\Big) - 2\kappa_\otimes\Big((x^+,x),(x'^+,x')\Big) \\ &\geq \kappa_\otimes\Big((x^+,x),(x^+,x)\Big) + \kappa_\otimes\Big((x'^+,x'),(x'^+,x')\Big) \\ &\quad - 2\gamma\sqrt{\kappa_\otimes\Big((x^+,x),(x^+,x)\Big)\kappa_\otimes\Big((x'^+,x'),(x'^+,x')\Big)} \\ &\geq 2B_\kappa'^2 - 2\gamma B_\kappa^2.\end{aligned} \tag{56}$$

That is, γ-coherent points in $\mathcal{D}_\gamma$ define two points in the image $\varphi(\mathbb{K}\times\mathbb{K})$ of the continuous feature map φ of the product kernel, i.e., $\left\{\varphi(x_i^+, x_i) : (x_i^+, x_i) \in \mathbb{K}\times\mathbb{K}\right\}$, that are at least $\sqrt{2B_\kappa'^2 - 2\gamma B_\kappa^2}$-separated. The image of a compact set under a continuous map is compact, and hence, $\varphi(\mathbb{K}\times\mathbb{K})$ is compact. Consequently, this set admits a finite covering and packing number[5] with balls of radius $\sqrt{2B_\kappa'^2 - 2\gamma B_\kappa^2}$. This packing number bounds $\mathcal{D}_\gamma$ from above. The logarithm of the packing number scales linearly with the dimension of the space, according to (Anthony et al., 1999). Then, the rest follows from $\dim(\mathbb{K}\times\mathbb{K}) = 2n$.

C. Proof of Theorem 3

Our derivation makes use of the following lemma based on Yu (1994, Corollary 2.7).

Lemma 2. *Kuznetsov & Mohri (2014, Proposition 2) Let $\{X_t\}_{t\in\mathbb{T}}$ be a β-mixing stochastic process on a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\in\mathbb{T}}, \mathbb{P})$ where X_t is $\mathcal{F}_t$-adapted and takes values on $\mathbb{X}\subseteq\mathbb{R}^n$ and Π a stationary distribution. Let $X^s(m)$ be an m-length sequence, sampled s time-points apart, and $X^\Pi(m)$ be an m-length sequence of i.i.d. samples drawn from Π. Then, a Borel measurable function $g : \mathbb{X}^m \to \mathbb{R}$ with $M_1 \le g(x) \le M_2$ for all $x \in \mathbb{X}^m$, satisfies*

$$\left|\mathbb{E}\left[g(X^\Pi(m))\right] - \mathbb{E}\left[g(X^s(m))\right]\right| \le (M_2 - M_1)m\beta(s). \tag{57}$$

We now utilize this lemma to establish the required result. Let $\widetilde{\mu}(X^\Pi(m))$ be the KME estimator constructed from an m-length i.i.d. sequence $X^\Pi(m) = (X_1^\Pi, \ldots, X_m^\Pi)$, given by

$$\widetilde{\mu}(X^\Pi(m)) = \frac{1}{m}\sum_{j=1}^{m}\kappa(X_j^\Pi, \cdot). \tag{58}$$

In the rest of the proof, we drop the dependency of X^s and X^Π on m for notational simplicity. For an m-length sequence Y, define a real-valued function $g : \mathbb{X}^m \to \{0, 1\}$ as

$$g(Y) := \mathbb{1}_{\{\|\widetilde{\mu}(Y)-\mu\|_\mathcal{H} - \mathbb{E}\|\widetilde{\mu}(X^\Pi)-\mu\|_\mathcal{H} > \epsilon\}}, \tag{59}$$

where $\mathbb{1}_{\{\cdot\}}$ denotes the indicator function. Lemma 2 applied to g gives

$$\mathbb{P}\left\{\|\widetilde{\mu}(X^s) - \mu\|_\mathcal{H} - \mathbb{E}\left\|\widetilde{\mu}(X^\Pi) - \mu\right\|_\mathcal{H} > \epsilon\right\} \le m\beta(s) + \mathbb{P}\left\{\left\|\widetilde{\mu}(X^\Pi) - \mu\right\|_\mathcal{H} - \mathbb{E}\left\|\widetilde{\mu}(X^\Pi) - \mu\right\|_\mathcal{H} > \epsilon\right\}. \tag{60}$$

We now bound the second term on the right hand side of the above inequality. To that end, notice that

$$\begin{aligned}\left\|\widetilde{\mu}(X^\Pi) - \mu\right\|_\mathcal{H} &= \sup_{\|f\|_\mathcal{H}\le 1}\left\langle f, \widetilde{\mu}(X^\Pi) - \mu\right\rangle \\ &= \sup_{\|f\|_\mathcal{H}\le 1}\left(\frac{1}{m}\sum_{i=1}^{m} f\left(X_i^\Pi\right) - \mathbb{E}[f(X)]\right),\end{aligned} \tag{61}$$

where $X \sim \Pi$. Under Assumption 1, we have

$$\begin{aligned}|f(x)|^2 &= \left|\langle f, \kappa(\cdot, x)\rangle_\mathcal{H}\right|^2 \\ &\le \|\kappa(\cdot, x)\|_\mathcal{H}^2\|f\|_\mathcal{H}^2 \\ &= \langle\kappa(\cdot, x), \kappa(\cdot, x)\rangle_\mathcal{H}\|f\|_\mathcal{H}^2 \\ &= \kappa(x, x)\|f\|_\mathcal{H}^2 \\ &\le B_\kappa\end{aligned} \tag{62}$$

for all $f \in \mathcal{H}$ with $\|f\|_\mathcal{H} \le 1$ and $x \in \mathbb{X}$, which implies

$$|f(x)| \le \sqrt{B_\kappa}. \tag{63}$$

[5] See (Anthony et al., 1999) for the definitions.

Now consider an m-length sequence $\widetilde{X}^{\Pi}$ that is identical to X^{Π}, except at the i-th position, where X_i^{Π} is replaced by an independently drawn sample from Π. The sequences X^{Π} and $\widetilde{X}^{\Pi}$, according to (61) and (63), satisfy

$$\left| \left\| \widetilde{\mu}(X^{\Pi}) - \mu \right\|_{\mathcal{H}} - \left\| \widetilde{\mu}(\widetilde{X}^{\Pi}) - \mu \right\|_{\mathcal{H}} \right| \leq \frac{2}{m} \sqrt{B_\kappa}. \tag{64}$$

This bounded difference property allows us to apply McDiarmid's inequality to infer

$$\begin{aligned} \mathbb{P}\left\{ \left\| \widetilde{\mu}(X^{\Pi}) - \mu \right\|_{\mathcal{H}} - \mathbb{E} \left\| \widetilde{\mu}(X^{\Pi}) - \mu \right\|_{\mathcal{H}} > \epsilon \right\} &\leq \exp\left(\frac{-2\epsilon^2}{m(2\sqrt{B_\kappa}/m)^2} \right) \\ &= \exp\left(\frac{-m\epsilon^2}{2B_\kappa} \right). \end{aligned} \tag{65}$$

Combining (60) and (65), we conclude

$$\mathbb{P}\left\{ \left\| \widetilde{\mu}(X^{s}) - \mu \right\|_{\mathcal{H}} - \mathbb{E} \left\| \widetilde{\mu}(X^{\Pi}) - \mu \right\|_{\mathcal{H}} > \epsilon \right\} \leq m\beta(s) + \exp\left(\frac{-m\epsilon^2}{2B_\kappa} \right).$$

In other words, we have

$$\left\| \widetilde{\mu}(X^{s}) - \mu \right\|_{\mathcal{H}} \leq \mathbb{E} \left\| \widetilde{\mu}(X^{\Pi}) - \mu \right\|_{\mathcal{H}} + \sqrt{\frac{2B_\kappa}{m} \log\left(\delta - m\beta(s) \right)^{-1}} \tag{66}$$

with probability at least $1 - \delta$ for any $\delta \in (m\beta(s), 1)$. The rest follows from bounding the first term on the right hand side of (66) as

$$\mathbb{E} \left\| \widetilde{\mu}(X^{\Pi}) - \mu \right\|_{\mathcal{H}} \leq \sqrt{B_\kappa / m}, \tag{67}$$

from Tolstikhin et al. (2017, Equation (47)).

D. Proof of Theorem 5

Assumption 3 guarantees that $\mathbb{E}[f(X_t)|X = \cdot] \in \mathcal{H}$ for all $f \in \mathcal{H}, t \in \mathbb{T}$, implying that $\mathcal{P}_t = C_t C_0^{-1}$. For a uniformly continuous semigroup on $\mathcal{H}$, its generator $\mathcal{A}$ becomes a bounded linear operator on $\mathcal{H}$, and thus, $\mathbb{D}(\mathcal{A}) = \mathcal{H}$, according to (Vrabie, 2003). We show uniform continuity in two steps. First, we prove that $\lim_{t \downarrow 0} \|C_t - C_0\|_{\text{HS}} = 0$, and second, we utilize that limit to argue uniform continuity.

• Proving $\lim_{t \downarrow 0} \|C_t - C_0\|_{\text{HS}} = 0$. Towards that goal, notice that

$$\begin{aligned} \|C_t - C_0\|_{\text{HS}} &= \Big\| \mathbb{E}\big[\phi(X_t) \otimes \phi(X_0)\big] - \mathbb{E}\big[\phi(X_0) \otimes \phi(X_0)\big] \Big\|_{\text{HS}} \\ &= \Big\| \mathbb{E}\Big[\mathbb{E}\big[\phi(X_t) \otimes \phi(X_0) | X_0\big] - \phi(X_0) \otimes \phi(X_0) \Big] \Big\|_{\text{HS}} \\ &\leq \mathbb{E}\Bigg[\underbrace{\Big\| \mathbb{E}\big[\phi(X_t) \otimes \phi(X_0) | X_0\big] - \phi(X_0) \otimes \phi(X_0) \Big\|_{\text{HS}}}_{\mathcal{T}_t} \Bigg], \end{aligned} \tag{68}$$

where the above derivation utilizes the tower property and Jensen's inequality. In the following, we show that $\lim_{t \downarrow 0} \mathcal{T}_t = 0$.

To that end, let X_t' be an independent copy of X_t given X_0. Then, we have

$$\begin{aligned}(\mathcal{T}_t)^2 =&\Big\|\mathbb{E}\big[\phi(X_t)\otimes\phi(X_0)|X_0\big]-\phi(X_0)\otimes\phi(X_0)\Big\|_{\text{HS}}^2\\ =&\Big\langle\mathbb{E}\big[\phi(X_t)\otimes\phi(X_0)|X_0\big]-\phi(X_0)\otimes\phi(X_0),\mathbb{E}\big[\phi(X_t')\otimes\phi(X_0)|X_0\big]-\phi(X_0)\otimes\phi(X_0)\Big\rangle_{\text{HS}}\\ =&\underbrace{\Big\langle\mathbb{E}\big[\phi(X_t)\otimes\phi(X_0)|X_0\big],\mathbb{E}\big[\phi(X_t')\otimes\phi(X_0)|X_0\big]\Big\rangle_{\text{HS}}}_{\mathcal{T}_t^{(1)}}\\ &-2\underbrace{\Big\langle\mathbb{E}\big[\phi(X_t)\otimes\phi(X_0)|X_0\big],\phi(X_0)\otimes\phi(X_0)\Big\rangle_{\text{HS}}}_{\mathcal{T}_t^{(2)}}\\ &+\underbrace{\Big\langle\phi(X_0)\otimes\phi(X_0),\phi(X_0)\otimes\phi(X_0)\Big\rangle_{\text{HS}}}_{\mathcal{T}_t^{(3)}},\end{aligned}\tag{69}$$

with the inner product in the Hilbert-Schmidt space defined in Gretton (2015). Denote the conditional expectation of X_t, X_t', and (X_t, X_t') given X_0 by $\mathop{\mathbb{E}}_{(X_t|X_0)}$, $\mathop{\mathbb{E}}_{(X_t'|X_0)}$, and $\mathop{\mathbb{E}}_{(X_t,X_t'|X_0)}$ respectively. Then, we have

$$\begin{aligned}\mathcal{T}_t^{(1)}&\overset{(a)}{=}\mathop{\mathbb{E}}_{(X_t|X_0)}\mathop{\mathbb{E}}_{(X_t'|X_0)}\Big[\big\langle\phi(X_t)\otimes\phi(X_0),\phi(X_t')\otimes\phi(X_0)\big\rangle_{\text{HS}}\Big]\\ &\overset{(b)}{=}\mathop{\mathbb{E}}_{(X_t|X_0)}\mathop{\mathbb{E}}_{(X_t'|X_0)}\Big[\big\langle\phi(X_t),\phi(X_t')\big\rangle_{\mathcal{H}}\big\langle\phi(X_0),\phi(X_0)\big\rangle_{\mathcal{H}}\Big]\\ &\overset{(c)}{=}\mathop{\mathbb{E}}_{(X_t|X_0)}\mathop{\mathbb{E}}_{(X_t'|X_0)}\Big[\kappa(X_t,X_t')\kappa(X_0,X_0)\Big]\\ &\overset{(d)}{=}\kappa(X_0,X_0)\mathop{\mathbb{E}}_{(X_t,X_t'|X_0)}\Big[\kappa(X_t,X_t')|X_0\Big],\end{aligned}\tag{70}$$

where (a) follows from the linearity of the (conditional) expectations and the inner product. Equality (b) follows from the definition of the inner product in $\mathcal{H}_\otimes$ and $\text{HS}(\mathcal{H})$. Equality (c) follows from the reproducing property of $\mathcal{H}$, and (d) follows from measurability of κ and the fact that X_t and X_t' are independent conditioned on X_0. Proceeding similarly with $\mathcal{T}_t^{(2)}$, we obtain

$$\begin{aligned}\mathcal{T}_t^{(2)}=&\mathop{\mathbb{E}}_{(X_t|X_0)}\Big[\big\langle\phi(X_t)\otimes\phi(X_0),\phi(X_0)\otimes\phi(X_0)\big\rangle_{\text{HS}}\Big]\\ =&\mathop{\mathbb{E}}_{(X_t|X_0)}\Big[\big\langle\phi(X_t),\phi(X_0)\big\rangle_{\mathcal{H}}\big\langle\phi(X_0),\phi(X_0)\big\rangle_{\mathcal{H}}\Big]\\ =&\kappa(X_0,X_0)\mathop{\mathbb{E}}_{(X_t|X_0)}\left[\kappa(X_t,X_0)|X_0\right],\end{aligned}\tag{71}$$

and $\mathcal{T}_t^{(3)}=\kappa^2(X_0,X_0)$. Combining the relationships obtained for $\mathcal{T}_t(i)$, $i=1,2,3$ in the definition of $\mathcal{T}_t$, we get

$$\mathcal{T}_t^2=\kappa(X_0,X_0)\mathop{\mathbb{E}}_{(X_t,X_t'|X_0)}\Big[\kappa(X_t,X_t')|X_0\Big]-2\kappa(X_0,X_0)\mathop{\mathbb{E}}_{(X_t|X_0)}\Big[\kappa(X_t,X_0)|X_0\Big]+\kappa^2(X_0,X_0).\tag{72}$$

Since $\{X_t\}_{t\in\mathbb{T}}$ is a diffusion process[6], its sample paths are almost surely continuous. Hence, boundedness and continuity of κ from Assumption 1 allow for the use of the dominated convergence theorem to infer almost surely,

$$\lim_{t\downarrow 0}\mathop{\mathbb{E}}_{(X_t|X_0)}\Big[\kappa(X_t,X_0)|X_0\Big]=\kappa(X_0,X_0),\quad\lim_{t\downarrow 0}\mathop{\mathbb{E}}_{(X_t,X_t'|X_0)}\Big[\kappa(X_t,X_t')|X_0\Big]=\kappa(X_0,X_0).\tag{73}$$

Utilizing the above relations in (72) implies $\lim_{t\downarrow 0}\mathcal{T}_t=0$, and therefore (68) yields

$$\lim_{t\downarrow 0}\|C_t-C_0\|_{\text{HS}}^2\le\lim_{t\downarrow 0}\left(\mathbb{E}\left[\mathcal{T}_t\right]\right)^2\le\lim_{t\downarrow 0}\mathbb{E}\left[(\mathcal{T}_t)^2\right]=0.\tag{74}$$

[6] in the sense of Schilling (2021, Definition 19.1).

The above derivation utilizes Jensen's inequality and the dominated convergence theorem. This concludes the proof of the first step.

• Concluding uniform continuity of $\{\mathcal{P}\}_{t\in\mathbb{T}}$. Using $\mathcal{P}_t = C_t C_0^{-1}$, we have

$$\begin{aligned}\lim_{t\downarrow 0}\|\mathcal{P}_t - \mathrm{id}\| &= \lim_{t\downarrow 0}\|\left(C_t - C_0\right)C_0^{-1}\| \\ &\leq \lim_{t\downarrow 0}\|C_t - C_0\|\|C_0^{-1}\| \\ &\leq \lim_{t\downarrow 0}\|C_t - C_0\|_{\mathrm{HS}}\|C_0^{-1}\| \\ &= 0.\end{aligned} \tag{75}$$

where we have used the sub-multiplicative nature of the operator norm and invertibility of C_0 from Assumption 3. This completes the proof of the step and the result.

E. Proof of Corollary 1

Under Assumption 3, we have $\mathcal{K}_t = C_0^{-1}C_t^\star$. Since $\|C_t - C_0\|_{\mathrm{HS}} = \|C_t^\star - C_0\|_{\mathrm{HS}}$, an argument identical to (74) yields that $\lim_{t\downarrow 0}\|\mathcal{K}_t - \mathrm{id}\| = 0$. It is easy to verify that $\{\mathcal{K}_t\}_{t\in\mathbb{T}}$ forms a semigroup, which in light of the above observation, becomes a uniformly continuous semigroup on $\mathcal{H}$. The rest follows from the fact that such semigroups admit bounded linear operators as generators whose domain is $\mathcal{H}$.

F. Proof of Theorem 6

We utilize the induction hypothesis to show that for any α with $|\alpha| < s$ and $f \in \mathcal{H}$, we have $D^\alpha\kappa(x,\cdot) \in \mathcal{H}$ and $\left(D^\alpha f\right)(x) = \langle D^\alpha\kappa(x,\cdot), f\rangle_{\mathcal{H}}$. The base case with $\alpha = 0$ holds trivially. Assume that the claim is true for some α with $|\alpha| < s$. We will prove the validity of the same for $(\alpha + e_j)$-order derivatives with $|\alpha + e_j| < s$ to complete the induction step. Here, the notation $\alpha + e_j$ stands for the vector of the order of partial differentiation, where the j-th coefficient is $\alpha_j + 1$. We prove the induction step in three parts.

First, we consider the function

$$\upsilon_x(t) := \frac{1}{t}\left[D^\alpha\kappa(x + te_j,\cdot) - D^\alpha\kappa(x,\cdot)\right] \tag{76}$$

for $t \in \mathbb{R}$ and show that for some sequence of t's converging to zero, $\upsilon_x(t)$ converges weakly to $D^{\alpha+e_j}\kappa(x,\cdot)$, using which we argue that $D^{\alpha+e_j}\kappa(x,\cdot)$ is an element of $\mathcal{H}$. Second, we show that $\upsilon_x(t)$ converges to $D^{\alpha+e_j}\kappa(x,\cdot)$ in $\mathcal{H}$-norm, and hence point-wise in $\mathcal{H}$. Finally, we utilize this pointwise convergence to demonstrate the reproducing property with $(\alpha + e_j)$-order derivatives.

• The induction hypothesis implies that $\upsilon_x(t) \in \mathcal{H}$ for all scalar $t \neq 0$. Also, the reproducing property of the induction hypothesis with $f = D^\alpha\kappa(y,\cdot)$ gives

$$\langle D^\alpha\kappa(x,\cdot), D^\alpha\kappa(y,\cdot)\rangle_{\mathcal{H}} = D^{(\alpha,\alpha)}\kappa(x,y). \tag{77}$$

Using the above relation and Taylor's expansion, we bound $\|\upsilon_x(t)\|_{\mathcal{H}}^2$ as

$$\begin{aligned}\|\upsilon_x(t)\|_{\mathcal{H}}^2 &= \frac{1}{t^2}\Big[D^{(\alpha,\alpha)}\kappa(x + te_j, x + te_j) - D^{(\alpha,\alpha)}\kappa(x + te_j, x) \\ &\quad - D^{(\alpha,\alpha)}\kappa(x, x + te_j) + D^{(\alpha,\alpha)}\kappa(x,x)\Big] \\ &\leq \|D^{(\alpha+e_j,\alpha+e_j)}\kappa\|_\infty\end{aligned} \tag{78}$$

The right-hand side of the inequality remains bounded for sufficiently small t, per Assumption 4. Thus, for small enough t's, $\upsilon_x(t)$'s lie in a closed ball of $\mathcal{H}$ with finite radius $\|D^{(\alpha+e_j,\alpha+e_j)}\kappa\|_\infty$. $\mathcal{H}$ is separable, and hence, by Banach-Alaoglu's Theorem (or alternatively, by Eberlein-Smulian Theorem), such a ball is weakly sequentially compact. Thus, for a sequence of t's going to zero, there is a subsequence (call it $\{t_k\}_{k=1}^\infty$) along which

$$\lim_{k\to\infty}\Big\langle\frac{1}{t_k}\left[D^\alpha\kappa(x + t_ke_j,\cdot) - D^\alpha\kappa(x,\cdot)\right], f\Big\rangle_{\mathcal{H}} = \Big\langle g_x, f\Big\rangle_{\mathcal{H}}, \quad \forall f \in \mathcal{H} \tag{79}$$

for some $g_x \in \mathcal{H}$. Plugging $f = \kappa(y, \cdot)$ for an arbitrary point $y \in \mathbb{R}^n$ in the above relation, we infer

$$\begin{aligned} g_x(y) &= \lim_{k\to\infty} \Big\langle \frac{1}{t_k}\big[D^\alpha \kappa(x + t_k e_j, \cdot) - D^\alpha \kappa(x, \cdot)\big], \kappa(y, \cdot)\Big\rangle_{\mathcal{H}} \\ &= \lim_{k\to\infty} \frac{1}{t_k}\big[D^\alpha \kappa(x + t_k e_j, y) - D^\alpha \kappa(x, y)\big] \\ &= D^{\alpha+e_j}\kappa(x, y), \end{aligned}$$

where we use the reproducing property of the induction hypothesis. Since $g_x \in \mathcal{H}$, we conclude that $D^{\alpha+e_j}\kappa(x, \cdot) \in \mathcal{H}$.

• Next, we show that $\upsilon_x(t)$ converges to the weak limit $D^{\alpha+e_j}\kappa(x, \cdot)$ in the $\mathcal{H}$-norm. With $f = D^{\alpha+e_j}\kappa(x, \cdot)$ in (79), the reproducing property of the induction hypothesis gives

$$\begin{aligned} \Big\|D^{\alpha+e_j}\kappa(x, \cdot)\Big\|^2_{\mathcal{H}} &= \Big\langle D^{\alpha+e_j}\kappa(x, \cdot), D^{\alpha+e_j}\kappa(x, \cdot)\Big\rangle_{\mathcal{H}} \\ &= \lim_{k\to\infty} \Big\langle \frac{1}{t_k}\big[D^\alpha \kappa(x + t_k e_j, \cdot) - D^\alpha \kappa(x, \cdot)\big], D^{\alpha+e_j}\kappa(x, \cdot)\Big\rangle_{\mathcal{H}} \\ &= \lim_{k\to\infty} \frac{1}{t_k}\Big[D^{(\alpha+e_j,\alpha)}\kappa(x, x + t_k e_j) - D^{(\alpha+e_j,\alpha)}\kappa(x, x)\Big] \\ &= D^{(\alpha+e_j,\alpha+e_j)}\kappa(x, x). \end{aligned}$$

Hence, using fundamental theorem of calculus, we infer

$$\begin{aligned} \Big\|\frac{1}{t}\big[D^\alpha \kappa(x + te_j, \cdot) - D^\alpha \kappa(x, \cdot)\big] - D^{\alpha+e_j}\kappa(x, \cdot)\Big\|^2_{\mathcal{H}} =& \frac{1}{t^2}\Big[D^{(\alpha,\alpha)}\kappa(x + te_j, x + te_j) - 2D^{(\alpha,\alpha)}\kappa(x + te_j, x) \\ &+ D^{(\alpha,\alpha)}\kappa(x, x)\Big] + D^{(\alpha+e_j,\alpha+e_j)}\kappa(x, x) \\ &- \frac{2}{t}\Big[D^{(\alpha+e_j,\alpha)}\kappa(x, x + te_j) - D^{(\alpha+e_j,\alpha)}\kappa(x, x)\big]\Big] \\ =& \frac{1}{t^2}\int_0^t \int_0^t D^{(\alpha+e_j,\alpha+e_j)}\kappa(x + ue_j, x + ve_j)\mathrm{d}u\mathrm{d}v \\ &- \frac{2}{t}\int_0^t D^{(\alpha+e_j,\alpha+e_j)}\kappa(x, x + ve_j)\mathrm{d}v \\ &+ D^{(\alpha+e_j,\alpha+e_j)}\kappa(x, x) \\ =& \frac{1}{t^2}\int_0^t \int_0^t \Big[D^{(\alpha+e_j,\alpha+e_j)}\kappa(x + ue_j, x + ve_j) \\ &- 2D^{(\alpha+e_j,\alpha+e_j)}\kappa(x, x + ve_j) + D^{(\alpha+e_j,\alpha+e_j)}\kappa(x, x)\Big]\mathrm{d}u\mathrm{d}v. \end{aligned}$$

Under Assumption 4, $D^{(\alpha+e_j,\alpha+e_j)}\kappa(x, x)$ is continuous and vanishes at infinity. Thus, $D^{(\alpha+e_j,\alpha+e_j)}\kappa(x, x)$ is uniformly continuous and the right hand side of the last line goes to 0 as $t \to 0$, which gives

$$\lim_{t\to 0}\Big\|\frac{1}{t}\big[D^\alpha \kappa(x + te_j, \cdot) - D^\alpha \kappa(x, \cdot)\big] - D^{\alpha+e_j}\kappa(x, \cdot)\Big\|^2_{\mathcal{H}} = 0.$$

Therefore, as $t \to 0$, $\frac{1}{t}\big[D^\alpha \kappa(x + te_j, \cdot) - D^\alpha \kappa(x, \cdot)\big]$ converges to $D^{\alpha+e_j}\kappa(x, \cdot)$ in $\mathcal{H}$-norm. Furthermore, since in RKHS, convergence in norm implies pointwise convergence (Steinwart & Christmann, 2008), we conclude that $\upsilon_x(t)$ converges pointwise to $D^\alpha \kappa(x, \cdot)$.

• Finally, we prove the partial derivative reproducing property of the induction step. For any $f \in \mathcal{H}$, we have

$$\begin{aligned} (D^{\alpha+e_j} f)(x) &:= \lim_{t\to 0} \frac{1}{t}\big[(D^\alpha f)(x + te_j) - (D^\alpha f)(x)\big] \\ &\overset{(a)}{=} \Big\langle \lim_{t\to 0} \frac{1}{t}\big[D^\alpha \kappa(x + te_j, \cdot) - D^\alpha \kappa(x, \cdot)\big], f\Big\rangle_{\mathcal{H}} \\ &\overset{(b)}{=} \Big\langle D^{\alpha+e_j}\kappa(x, \cdot), f\Big\rangle_{\mathcal{H}}, \end{aligned}$$

where (a) holds due to the induction hypothesis and (b) follows from point-wise convergence proved in previous step. This completes the proof.

G. Proof of Lemma 1

Let X, X' be independently and identically distributed with distribution $\mathbb{P}_X$. The HS-norm of ∂C_0 can be computed as

$$
\begin{aligned}
\|\partial C_0\|_{\text{HS}}^2 =& \|\mathbb{E}[\mathrm{d}^{(2)}\phi(X)\otimes\phi(X)]\|_{\text{HS}}^2 \\
=& \left|\left\langle \mathbb{E}[\mathrm{d}^{(2)}\phi(X)\otimes\phi(X)], \mathbb{E}[\mathrm{d}^{(2)}\phi(X)\otimes\phi(X)]\right\rangle_{\text{HS}}\right| \\
\leq& \mathbb{E}\left|\left\langle \mathrm{d}^{(2)}\phi(X)\otimes\phi(X), \mathrm{d}^{(2)}\phi(X')\otimes\phi(X')\right\rangle_{\text{HS}}\right| \\
=& \mathbb{E}\left|\left\langle \mathrm{d}^{(2)}\phi(X), \mathrm{d}^{(2)}\phi(X')\right\rangle_{\mathcal{H}} \langle\phi(X),\phi(X')\rangle_{\mathcal{H}}\right| \\
\leq& \mathbb{E}\left[\left\|\mathrm{d}^{(2)}\phi(X)\right\|_{\mathcal{H}} \left\|\mathrm{d}^{(2)}\phi(X')\right\|_{\mathcal{H}} \|\phi(X)\|_{\mathcal{H}} \|\phi(X')\|_{\mathcal{H}}\right] \\
=& \mathbb{E}\left[\left\|\mathrm{d}^{(2)}\phi(X)\right\|_{\mathcal{H}}^2 \|\phi(X)\|_{\mathcal{H}}^2\right] \\
=& \mathbb{E}\left[\left(\sum_i^n\sum_j^n b_i(X)b_j(X)D^{(e_i,e_j)}\kappa(X,X)\right.\right. \\
&+\frac{1}{2}\sum_i^n\sum_p^n\sum_q^n b_i(X)a_{pq}(X)D^{(e_i,e_p+e_q)}\kappa(X,X) \\
&+\frac{1}{2}\sum_i^n\sum_j^n\sum_q^n a_{pq}(X)b_i(X)D^{(e_i+e_j,e_p)}\kappa(X,X) \\
&\left.\left.+\frac{1}{4}\sum_i^n\sum_j^n\sum_p^n\sum_q^n a_{ij}(X)a_{pq}(x_s)D^{(e_i+e_j,e_p+e_q)}\kappa(X,X)\right)\kappa(X,X)\right] \\
\leq& B_\kappa^2\left(n^2+n^3+\frac{1}{4}n^4\right)\overline{S}.
\end{aligned}
\tag{80}
$$

Therefore, we conclude that ∂C_0 defined in (29) is a Hilbert Schmidt operator.

H. Proof of Theorem 7

The proof follows from the same argument as that of Theorem 1 . In particular, replacing $C_{X^+X}, \widehat{C}_{X^+X}$ with $\partial C_0, \widehat{\partial C}_0$ in Equations (39), (40), (42) and (43) gives

$$
\begin{aligned}
\left\|\widetilde{\mathcal{A}}_\varepsilon - \mathcal{A}_\varepsilon\right\| \leq& \frac{1}{\varepsilon}\left\|\widehat{\partial C}_0 - \partial C_0\right\|_{\text{HS}} + \frac{1}{\varepsilon^2}\left\|C_0-\widetilde{C}_0\right\|_{\text{HS}}\|\partial C_0\| \\
\overset{(a)}{\leq}& \frac{1}{\varepsilon}\left(\sqrt{\frac{B_\otimes}{M}}\Xi(\delta/2)\right) + \frac{1}{\varepsilon^2}\frac{B_\kappa}{\sqrt{M}}\Xi(\delta/2)\|\partial C_0\|_{\text{HS}} \\
\leq& \frac{1}{\sqrt{M}}\Xi(\delta/2)\left(\frac{1}{\varepsilon}\sqrt{B_\otimes} + \frac{\|\partial C_0\|_{\text{HS}}}{\varepsilon^2}B_\kappa\right) \\
\leq& \frac{1}{\sqrt{M}}\Xi(\delta/2)\left(\frac{1}{\varepsilon}+\frac{\|\partial C_0\|_{\text{HS}}}{\varepsilon^2}\right)\max\left(\sqrt{B_\otimes}, B_\kappa\right) \\
=& \frac{1}{\sqrt{M}}\Xi(\delta/2)B^{\max}\mathcal{O}(\varepsilon^{-2}),
\end{aligned}
\tag{81}
$$

where (a) holds with probability at least $1-\delta$.

I. Proof of Theorem 8

Let $\mathcal{I}$ denotes the indices among $1, \cdots, M$ for which $x_k \in \mathcal{D}_\gamma$. Denote $r(x) := \mathrm{d}^{(2)}\phi(x) \otimes \phi(x)$. Under condition (33), $\mathcal{D}_\gamma$ satisfies for any $q \in \mathcal{I}$,

$$\begin{aligned}
\max_{p\in\mathcal{I},p\neq q} \frac{\left|\langle r(x_p), r(x_q)\rangle_{\mathcal{H}_\otimes}\right|}{\|r(x_p)\|_{\mathcal{H}_\otimes}\|r(x_q)\|_{\mathcal{H}_\otimes}} &= \max_{p\in\mathcal{I},p\neq q} \frac{\left|\left\langle \mathrm{d}^{(2)}\phi\left(x_p\right)\otimes\phi\left(x_p\right), \mathrm{d}^{(2)}\phi\left(x_q\right)\otimes\phi\left(x_q\right)\right\rangle_{\mathcal{H}_\otimes}\right|}{\left\|\mathrm{d}^{(2)}\phi\left(x_p\right)\otimes\phi\left(x_p\right)\right\|_{\mathcal{H}_\otimes}\left\|\mathrm{d}^{(2)}\phi\left(x_q\right)\otimes\phi\left(x_q\right)\right\|_{\mathcal{H}_\otimes}} \\
&= \max_{p\in\mathcal{I},p\neq q} \frac{\left|\left\langle \mathrm{d}^{(2)}\phi\left(x_p\right), \mathrm{d}^{(2)}\phi\left(x_q\right)\right\rangle_{\mathcal{H}}\left\langle \phi\left(x_p\right), \phi\left(x_q\right)\right\rangle_{\mathcal{H}}\right|}{\left\|\mathrm{d}^{(2)}\phi\left(x_p\right)\right\|_{\mathcal{H}}\left\|\phi\left(x_p\right)\right\|_{\mathcal{H}}\left\|\mathrm{d}^{(2)}\phi\left(x_q\right)\right\|_{\mathcal{H}}\left\|\phi\left(x_q\right)\right\|_{\mathcal{H}}} \\
&= \max_{p\in\mathcal{I},p\neq q} \frac{\left|\left\langle \mathrm{d}^{(2)}\phi\left(x_p\right), \mathrm{d}^{(2)}\phi\left(x_q\right)\right\rangle_{\mathcal{H}}\right|}{\left\|\mathrm{d}^{(2)}\phi\left(x_p\right)\right\|_{\mathcal{H}}\left\|\mathrm{d}^{(2)}\phi\left(x_q\right)\right\|_{\mathcal{H}}} \frac{\left|\left\langle \phi\left(x_p\right), \phi\left(x_q\right)\right\rangle_{\mathcal{H}}\right|}{\left\|\phi\left(x_p\right)\right\|_{\mathcal{H}}\left\|\phi\left(x_q\right)\right\|_{\mathcal{H}}} \\
&\leq \gamma.
\end{aligned} \tag{82}$$

Thus we have

$$\max_{p\in\mathcal{I},p\neq q} \frac{\left|\langle r(x_p), r(x_q)\rangle_{\mathcal{H}_\otimes}\right|}{\|r(x_p)\|_{\mathcal{H}_\otimes}} \leq \gamma\|r(x_q)\|_{\mathcal{H}_\otimes} \leq \gamma\sqrt{B_\otimes}, \tag{83}$$

where the bound on $r(\cdot)$ follows from the intermediate step in (80).

Let $\Pi_{\mathcal{D}_\gamma}$ denote the projection operator onto the subspace spanned by a dictionary of functions $\{r(x_k) : x_k \in \mathcal{D}_\gamma\}$. The error due to sparsification can be upper bounded by

$$\begin{aligned}
\|\widetilde{\partial C}_0 - \widehat{\partial C}_0\|_{\mathrm{HS}} &= \|\frac{1}{M}\sum_{m=1}^{M}(\mathrm{id} - \Pi_{\mathcal{D}_\gamma})r(x_m)\|_{\mathcal{H}_\otimes} \\
&\leq \sum_{m=1}^{M}\frac{1}{M}\|r(x_m) - \Pi_{\mathcal{D}_\gamma}r(x_m)\|_{\mathcal{H}_\otimes} \\
&= \frac{1}{M}\sum_{m\notin\mathcal{I}}\|r(x_m) - \Pi_{\mathcal{D}_\gamma}r(x_m)\|_{\mathcal{H}_\otimes} \\
&\leq \left(1 - \frac{|\mathcal{D}_\gamma|}{M}\right)\sqrt{B_\otimes(1-\gamma^2)}.
\end{aligned} \tag{84}$$

In the following, we justify the last inequality in (84). For any $x_p \notin \mathcal{D}_\gamma$, we have

$$\begin{aligned}
\|\Pi_{\mathcal{D}_\gamma}r(x_p)\|_{\mathcal{H}_\otimes} &= \max_{\alpha}\frac{\langle\sum_{s\in\mathcal{I}}\alpha_s r(x_s), r(x_p)\rangle_{\mathcal{H}_\otimes}}{\|\sum_{s\in\mathcal{I}}\alpha_s r(x_s)\|_{\mathcal{H}_\otimes}} \\
&\geq \max_{k\in\mathcal{I}}\frac{\langle r(x_k), r(x_p)\rangle_{\mathcal{H}_\otimes}}{\|r(x_k)\|_{\mathcal{H}_\otimes}} \\
&\geq \gamma\sqrt{B_\otimes},
\end{aligned} \tag{85}$$

where the first inequality follows from a specific choice of coefficients where $\alpha_j = 0$ for each $j \in \mathcal{I}$, except for a single index q with $\alpha_k = \pm 1$. The second inequality follows from the violation of (83) since $x_p \notin \mathcal{D}_\gamma$. Hence, for $p \notin \mathcal{I}$, Pythagoras theorem gives

$$\begin{aligned}
\|r(x_p) - \Pi_{\mathcal{D}_\gamma}r(x_p)\|^2_{\mathcal{H}_\otimes} &= \|r(x_p)\|^2_{\mathcal{H}_\otimes} - \|\Pi_{\mathcal{D}_\gamma}r(x_p)\|^2_{\mathcal{H}_\otimes} \\
&\leq B_\otimes(1-\gamma^2),
\end{aligned} \tag{86}$$

This justifies (84).

On the other hand, using Theorem 7, we have that with probability at least $1-\delta$,

$$\left\|\partial C_0 - \widetilde{\partial C}_0\right\|_{HS} \leq \sqrt{\frac{B_\otimes}{M}}\Xi(\delta). \tag{87}$$

Therefore, along the same lines as (43) in Appendix A, we have with probability at least $1-\delta$

$$\begin{aligned}\|\widehat{\partial C}_0 - \partial C_0\|_{\text{HS}} &\leq \|\widehat{\partial C}_0 - \widetilde{\partial C}_0\|_{\text{HS}} + \|\widetilde{\partial C}_0 - \partial C_0\|_{\text{HS}} \\ &\leq \left(1 - \frac{|\mathcal{D}_\gamma|}{M}\right)\sqrt{B_\otimes(1-\gamma^2)} + \sqrt{\frac{B_\otimes}{M}}\Xi(\delta),\end{aligned} \tag{88}$$

where the second inequality follows from (84) and (87). This implies

$$\begin{aligned}\left\|\widehat{\mathcal{A}}_\varepsilon - \mathcal{A}_\varepsilon\right\| \leq& \frac{1}{\varepsilon}\left\|\widehat{\partial C}_0 - \partial C_0\right\|_{\text{HS}} + \frac{1}{\varepsilon^2}\left\|C_0 - \widehat{C}_0\right\|_{\text{HS}}\|\partial C_0\| \\ \leq& \frac{1}{\varepsilon}\left(\left(1 - \frac{|\mathcal{D}_\gamma|}{M}\right)\sqrt{B_\otimes(1-\gamma^2)} + \sqrt{\frac{B_\otimes}{M}}\Xi(\delta/2)\right) \\ &+ \frac{1}{\varepsilon^2}\left(\left(1 - \frac{|\mathcal{D}_\gamma|}{M}\right)B_\kappa\sqrt{1-\gamma^2} + \frac{B_\kappa}{\sqrt{M}}\Xi(\delta/2)\right)\|\partial C_0\| \\ \leq& \left(1 - \frac{|\mathcal{D}_\gamma|}{M}\right)\left(\frac{1}{\varepsilon}\sqrt{B_\otimes(1-\gamma^2)} + \frac{\|\partial C_0\|}{\varepsilon^2}B_\kappa\sqrt{1-\gamma^2}\right) \\ &+ \frac{1}{\sqrt{M}}\Xi(\delta/2)\left(\frac{1}{\varepsilon}\sqrt{B_\otimes} + \frac{\|\partial C_0\|}{\varepsilon^2}B_\kappa\right) \\ \leq& \left(1 - \frac{|\mathcal{D}_\gamma|}{M}\right)\left(\frac{1}{\varepsilon} + \frac{\|\partial C_0\|}{\varepsilon^2}\right)\sqrt{1-\gamma^2}\max\left(\sqrt{B_\otimes}, B_\kappa\right) \\ &+ \frac{1}{\sqrt{M}}\Xi(\delta/2)\left(\frac{1}{\varepsilon} + \frac{\|\partial C_0\|}{\varepsilon^2}\right)\max\left(\sqrt{B_\otimes}, B_\kappa\right) \\ =& B^{\max}\left(\left(1 - \frac{|\mathcal{D}_\gamma|}{M}\right)\sqrt{1-\gamma^2} + \frac{1}{\sqrt{M}}\Xi(\delta/2)\right)\mathcal{O}(\varepsilon^{-2}).\end{aligned} \tag{89}$$

J. Conditional Joint Mean Embedding

We discuss the notion of conditional joint mean embedding (also known as conditional cross covariance in Song et al. (2009)). Consider an RKHS $\mathcal{H}$ with kernel κ defined over some subset $\mathbb{X}$ of an Euclidean space, where ϕ is the feature map. Then, the cross product $\phi(X)\otimes\phi(Y)\otimes\phi(Z)$ is a random variable in $\mathcal{H}_\otimes^\circ := \mathcal{H}\otimes\mathcal{H}\otimes\mathcal{H}$. When $\mathbb{E}\left[\|\phi(X)\otimes\phi(Y)\otimes\phi(Z)\|_{\mathcal{H}_\otimes^\circ}\right] < \infty$, we define the joint covariance operator $C_{XYZ} : \mathcal{H}_\otimes \to \mathcal{H}$ as

$$C_{XYZ} := \mathbb{E}\left[\phi(X)\otimes\phi(Y)\otimes\phi(Z)\right]. \tag{90}$$

Thus for $f, g, r \in \mathcal{H}$, we have

$$\begin{aligned}\mathbb{E}\left[f(X)g(Y)r(Z)\right] &= \mathbb{E}\left[\langle\phi(X), f\rangle_{\mathcal{H}}\langle\phi(Y), g\rangle_{\mathcal{H}}\langle\phi(Z), r\rangle_{\mathcal{H}}\right] \\ &= \mathbb{E}\left[\langle\phi(X), f\rangle_{\mathcal{H}}\langle\phi(Y)\otimes\phi(Z), g\otimes r\rangle_{\mathcal{H}_\otimes}\right] \\ &= \mathbb{E}\left[\left\langle\phi(X)\otimes\big(\phi(Y)\otimes\phi(Z)\big), f\otimes\big(g\otimes r\big)\right\rangle_{\mathcal{H}_\otimes^\circ}\right] \\ &= \left\langle\mathbb{E}\left[\phi(X)\otimes\phi(Y)\otimes\phi(Z)\right], f\otimes g\otimes r\right\rangle_{\mathcal{H}_\otimes^\circ} \\ &= \langle C_{XYZ}, f\otimes g\otimes r\rangle.\end{aligned} \tag{91}$$

Also, the above equals $\langle f, C_{XYZ}(g\otimes r)\rangle$. Given M data points $\mathcal{M} := \{(x_m, y_m, z_m)\}_{m=1}^M$ sampled i.i.d. from the joint distribution of X, Y, Z, the empirical estimate of C_{XYZ} is given by

$$\widetilde{C}_{XYZ} = \frac{1}{M}\sum_{m=1}^M \phi(x_m)\otimes\phi(y_m)\otimes\phi(z_m). \tag{92}$$

We now define the conditional joint mean embedding operator $\mathcal{U}_{ZY|X}$. Assume

$$\mathbb{E}\left[g(Y)r(Z)|X=\cdot\right]\in\mathcal{H},\quad\forall g,r\in\mathcal{H}. \tag{93}$$

Similar to Fukumizu et al. (2004, Theorem 2), we have

$$C_{XX}\mathbb{E}\left[g\left(Y\right)r\left(Z\right)|X=\cdot\right]=C_{XYZ}\left(g\otimes r\right). \tag{94}$$

Together with the reproducing property, we obtain

$$\begin{aligned}\mathbb{E}\left[g(Y)r(Z)|X=x\cdot\right]=&\left\langle\mathbb{E}\left[g(Y)r(Z)|X\cdot\right],\kappa(x,\cdot)\right\rangle_{\mathcal{H}}\\=&\left\langle C_{XX}^{-1}C_{XYZ}\left(g\otimes r\right),\kappa(x,\cdot)\right\rangle_{\mathcal{H}}\\=&\left\langle g\otimes r,C_{ZYX}C_{XX}^{-1}\kappa(x,\cdot)\right\rangle_{\mathcal{H}_{\otimes}}.\end{aligned} \tag{95}$$

Therefore, the operator $\mathcal{U}_{ZY|X}$ and mapping $\mu_{ZY|x}$ can be defined as

$$\mathcal{U}_{ZY|X}:=C_{ZYX}C_{XX}^{-1},\quad\mu_{ZY|x}:=\mathbb{E}\left[\phi(Y)\otimes\phi(Z)|X=x\right]=\mathcal{U}_{ZY|X}\kappa(x,\cdot), \tag{96}$$

such that

$$\mathbb{E}\left[g(Y)r(Z)|X=x\right]=\left\langle g\otimes r,\mu_{ZY|x}\right\rangle_{\mathcal{H}_{\otimes}} \tag{97}$$

In other words, $\mathcal{U}_{tt'}$ and $\mathcal{U}_{\tau 0}$ in Section 8 can be defined as

$$\mathcal{U}_{tt'}:=C_{tt'0}C_0^{-1}. \tag{98}$$

One can also define the finite sample estimations of these operators with regularization parameter ε as

$$\widehat{\mathcal{U}}_{tt'}^{\varepsilon}:=\widehat{C}_{tt'0}(\widehat{C}_0+\varepsilon\,\mathrm{id})^{-1}. \tag{99}$$

The empirical estimates of b^{τ},a^{τ} can be computed using these data-driven variants. In addition, the evaluation at a data point can be constructed using finite dimension Gram matrices as in Appendix K.

K. Estimate Drift and Diffusion Coefficients Using Finite Dimensional Matrices

In practice often drift and diffusion terms are not known. In the following, we show that for implementing the algorithm proposed in Section 8, Gram matrices are only required. To simplify our notations, we consider a 1-dimensional dynamical system. The n-dimensional extension can be derived similarly.

Let $\{(x_i\left(0\right),x_i\left(\tau\right))\}_{i=1}^{N}$ be N snapshot pairs used to construct empirical estimates of $\widehat{\mathcal{U}}_{\tau}^{\varepsilon}$, $\widehat{\mathcal{U}}_{\tau\tau}^{\varepsilon}$ and $\widehat{\mathcal{U}}_{\tau 0}^{\varepsilon}$ in (99). Define feature matrices

$$\begin{aligned}\Phi_{\tau}&=\left[\phi\left(x_1\left(\tau\right)\right),\ldots,\phi\left(x_N\left(\tau\right)\right)\right],\\\Phi_0&=\left[\phi\left(x_1\left(0\right)\right),\ldots,\phi\left(x_N\left(0\right)\right)\right],\end{aligned}$$

and Gram matrix $G_{00}^{\circ} := \Phi_0^{\top}\Phi_0$. Then, for $z \in \mathbb{R}^n$,

$$\begin{aligned}\left[\left(\widehat{\mathcal{U}}_{\tau}^{\varepsilon}\right)^{\star}\mathrm{id}\right](z) = \left\langle\left(\widehat{\mathcal{U}}_{\tau}^{\varepsilon}\right)^{\star}\mathrm{id}, \kappa(z,\cdot)\right\rangle &= \left\langle \mathrm{id}, \widehat{\mathcal{U}}_{\tau}^{\varepsilon}\kappa(z,\cdot)\right\rangle \\ &= \left\langle \mathrm{id}, \widehat{C}_{\tau}\left(\widehat{C}_0 + \varepsilon\,\mathrm{id}\right)^{-1}\kappa(z,\cdot)\right\rangle \\ &\overset{(a)}{=} \left\langle \mathrm{id}, \Phi_{\tau}\underbrace{\left(G_{00}^{\circ} + \varepsilon\,\mathrm{id}\right)^{-1}\Phi_0^{\top}\kappa(z,\cdot)}_{:=g(z)}\right\rangle \\ &= \left\langle \mathrm{id}, \sum_{i=1}^{N} g_i(z)\,\kappa\left(x_i(\tau),\cdot\right)\right\rangle \\ &= \sum_{i=1}^{N} g_i(z)\left\langle \mathrm{id}, \kappa\left(x_i(\tau),\cdot\right)\right\rangle \\ &= \sum_{i=1}^{N} g_i(z)\,x_i(\tau)\,.\end{aligned} \tag{100}$$

where (a) follows from relation (106) and $\left(G_{00}^{\circ} + \varepsilon\,\mathrm{id}\right)^{-1}\Phi_0^{\top}\kappa(z,\cdot) := g(z) \in \mathbb{R}^N$.

Therefore, we have

$$\begin{aligned}\widehat{b^{\tau}}(z) &= \frac{1}{\tau}\left\{\left[\left(\widehat{\mathcal{U}}_{\tau}^{\varepsilon}\right)^{\star}\mathrm{id}\right](z) - (\mathrm{id})(z)\right\} \\ &= \frac{1}{\tau}\left\{\sum_{i=1}^{N} g_i(z)\,x_i(\tau) - z\right\}.\end{aligned} \tag{101}$$

In the above calculation, the first line requires id to be an element in $\mathcal{H}$. As we noted in Section 8, the identity map does not belong to an infinite-dimensional RKHS. However, the calculation utilizes the action of the identity map, and not its representation within $\mathcal{H}$. In other words, the above derivation can be viewed as a formal approximation technique. Providing guarantees on approximation error, however, remains challenging. In order to construct $\widehat{a^{\tau}}$, define a feature matrix $\Phi_{\tau\tau}$ whose i-th column is $\phi\left(x_i(\tau)\right) \otimes \phi\left(x_i(\tau)\right)$. Using (92), we have

$$\begin{aligned}\left[\left(\widehat{\mathcal{U}}_{\tau\tau}^{\varepsilon}\right)^{\star}\right](\mathrm{id}\otimes\mathrm{id})(z) &= \left\langle \mathrm{id}\otimes\mathrm{id}, \widehat{\mathcal{U}}_{\tau\tau}^{\varepsilon}\kappa(z,\cdot)\right\rangle \\ &= \left\langle \mathrm{id}, \widehat{C}_{\tau\tau 0}\left(\widehat{C}_0 + \varepsilon\,\mathrm{id}\right)^{-1}\kappa(z,\cdot)\right\rangle \\ &= \left\langle \mathrm{id}, \Phi_{\tau\tau}\underbrace{\left(G_{00}^{\circ} + \varepsilon\,\mathrm{id}\right)^{-1}\Phi_0^{\top}\kappa(z,\cdot)}_{:=g(z)}\right\rangle \\ &= \left\langle \mathrm{id}\otimes\mathrm{id}, \sum_{i=1}^{N} g_i(z)\,\kappa\left(x_i(\tau),\cdot\right)\otimes\kappa\left(x_i(\tau),\cdot\right)\right\rangle \\ &= \sum_{i=1}^{N} g_i(z)\left\langle \mathrm{id}\otimes\mathrm{id}, \kappa\left(x_i(\tau),\cdot\right)\otimes\kappa\left(x_i(\tau),\cdot\right)\right\rangle \\ &= \sum_{i=1}^{N} g_i(z)\,x_i^2(\tau)\,.\end{aligned} \tag{102}$$

Thus, $\widehat{a^{\tau}}(z)$ can also be computed using Gram matrices as

$$\widehat{a^{\tau}}(z) = \frac{1}{\tau}\left\{\sum_{i=1}^{N} g_i(z)\,x_i^2(\tau) - 2\sum_{i=1}^{N} g_i(z)\,x_i(\tau)\,x_i(0) + z^2\right\}. \tag{103}$$

Again, we utilize the action of id, without its explicit representation in $\mathcal{H}$.

L. Algorithm Summary

Algorithm 1 Sample-based Sparse Learning of $\mathcal{A}$ Without Explicit Knowledge of SDE Coefficients

1: **Input:** Sampling interval τ; snapshot pairs $\mathcal{D}_1 := \{x_m(0), x_m(\tau)\}_{m=1}^N$; second dataset $\mathcal{D}_2 := \{x_m\}_{m=1}^M$; kernel function κ; coherence parameters γ_1, γ_2; data dimension n.

2: **Output:** Sparse, regularized estimator $\widehat{\mathcal{A}}_\varepsilon^\tau$.

3: Prune $\mathcal{D}_1$ to get $\mathcal{D}_{\gamma_1}$ by identifying a subset of $\mathcal{D}_1$ such that each pair $(x_i(0), x_i(\tau)), (x_j(0), x_j(\tau)) \in \mathcal{D}_{\gamma_1}$ satisfies

$$|\kappa(x_i(0), x_j(0))| \le \sqrt{\gamma_1 \kappa(x_i(0), x_i(0))\kappa(x_j(0), x_j(0))}, \quad |\kappa(x_i(\tau), x_j(\tau))| \le \sqrt{\gamma_1 \kappa(x_i(\tau), x_i(\tau))\kappa(x_j(\tau), x_j(\tau))}.$$

4: Let $\mathcal{I}_1$ be the indices among $1, \dots, N$ for which $(x_i(0), x_i(\tau))$ are in $\mathcal{D}_{\gamma_1}$. Construct

$$\widehat{C}_{\tau 0} = \sum_{i \in \mathcal{I}_1} \alpha_i \varphi\left(x_i(\tau), x_i(0)\right), \quad \widehat{C}_0 = \sum_{i \in \mathcal{I}_1} \beta_i \varphi\left(x_i(0), x_i(0)\right),$$

where α (and similarly, β) [7] is defined via $\alpha = G^{-1} g$, with $G_{i,j} = \kappa_\otimes((x_i(\tau), x_i(0)), (x_j(\tau), x_j(0)))$ for $i, j \in \mathcal{I}_1$, and $g_j = \frac{1}{N}\sum_{i=1}^N \kappa_\otimes((x_i(\tau), x_i(0)), (x_j(\tau), x_j(0)))$ for $j \in \mathcal{I}_1$. $\widehat{C}_{\tau\tau 0}$ and $\widehat{C}_{\tau 00}$ can be computed likewise.

5: Compute $\widehat{\mathcal{U}}_\tau^\varepsilon, \widehat{\mathcal{U}}_{\tau\tau}^\varepsilon$, and $\widehat{\mathcal{U}}_{\tau 0}^\varepsilon$ based on $\mathcal{D}_{\gamma_1}$ via

$$\widehat{\mathcal{U}}_\tau^\varepsilon := \widehat{C}_{\tau 0}(\widehat{C}_0 + \varepsilon \,\mathrm{id})^{-1}, \quad \widehat{\mathcal{U}}_{\tau\tau}^\varepsilon := \widehat{C}_{\tau\tau 0}(\widehat{C}_0 + \varepsilon \,\mathrm{id})^{-1}, \quad \widehat{\mathcal{U}}_{\tau 0}^\varepsilon := \widehat{C}_{\tau 00}(\widehat{C}_0 + \varepsilon \,\mathrm{id})^{-1}.$$

6: Construct sparse estimates of $\widehat{b^\tau}, \widehat{a^\tau}$ based on $\mathcal{D}_{\gamma_1}$ via (see Appendix K)

$$\begin{aligned}
\widehat{b_i^\tau} =& \frac{1}{\tau}\left(\left(\widehat{\mathcal{U}}_\tau^\varepsilon\right)^\star \boldsymbol{e}_i - \boldsymbol{e}_i\right), \quad \forall i = 1, \cdots, n, \\
\widehat{a_{ij}^\tau} =& \frac{1}{\tau}\left(\left(\widehat{\mathcal{U}}_{\tau\tau}^\varepsilon\right)^\star (\boldsymbol{e}_i \otimes \boldsymbol{e}_j) - \left(\widehat{\mathcal{U}}_{\tau 0}^\varepsilon\right)^\star (\boldsymbol{e}_i \otimes \boldsymbol{e}_j) - \left(\widehat{\mathcal{U}}_{\tau 0}^\varepsilon\right)^\star (\boldsymbol{e}_j \otimes \boldsymbol{e}_i) + (\boldsymbol{e}_i \otimes \boldsymbol{e}_j)\kappa(x, \cdot)\right), \ \forall i, j = 1, \cdots, n.
\end{aligned}$$

7: Compute $\widehat{\mathrm{d}^{(2)}\phi}$ via $\widehat{\mathrm{d}^{(2)}\phi} = \sum_{i=1}^n \widehat{b_i^\tau} D^{e_i}\phi + \frac{1}{2}\sum_{i=1}^n \sum_{j=1}^n \widehat{a_{ij}^\tau} D^{e_i + e_j}\phi$. Then prune $\mathcal{D}_2$ to get $\mathcal{D}_{\gamma_2}$ by identifying a subset of $\mathcal{D}_2$ such that each pair $x_p, x_q \in \mathcal{D}_{\gamma_2}$ satisfies

$$\langle \widehat{\mathrm{d}^{(2)}\phi}(x_p), \widehat{\mathrm{d}^{(2)}\phi}(x_q)\rangle_{\mathcal{H}} \le \sqrt{\gamma_2}\|\widehat{\mathrm{d}^{(2)}\phi}(x_p)\|_{\mathcal{H}}\|\widehat{\mathrm{d}^{(2)}\phi}(x_q)\|_{\mathcal{H}}, \quad |\kappa(x_p, x_q)| \le \sqrt{\gamma_2 \kappa(x_p, x_q)\kappa(x_p, x_q)}$$

8: Let $\mathcal{I}_2$ be the indices among $1, \dots, M$ for which $x_k \in \mathcal{D}_{\gamma_2}$. Compute $\widehat{\partial C}_0(\mathcal{D}_{\gamma_1}, \mathcal{D}_{\gamma_2})$, $\widehat{C}_0(\mathcal{D}_{\gamma_2})$ via

$$\widehat{\partial C}_0\left(\mathcal{D}_{\gamma_1}, \mathcal{D}_{\gamma_2}\right) := \sum_{k \in \mathcal{I}_2} z_k \, \widehat{\mathrm{d}^{(2)}\phi}(x_k) \otimes \phi(x_k), \quad \widehat{C}_0\left(\mathcal{D}_2\right) = \sum_{k \in \mathcal{I}_2} z_k' \varphi\left(x_k, x_k\right),$$

where z (and similarly, z')[7] is obtained as $z = H^{-1}h$, with $H_{i,j} = \langle \mathrm{d}^{(2)}\phi(x_i) \otimes \phi(x_i), \mathrm{d}^{(2)}\phi(x_j) \otimes \phi(x_j)\rangle$ for all$i, j \in \mathcal{I}_2$,$h_j = \frac{1}{M}\sum_{i=1}^M \langle \mathrm{d}^{(2)}\phi(x_i) \otimes \phi(x_i), \mathrm{d}^{(2)}\phi(x_j) \otimes \phi(x_j)\rangle, \quad \forall j \in \mathcal{I}_2$.

9: Compute $\widehat{\mathcal{A}}_\varepsilon^\tau(\mathcal{D}_{\gamma_1}, \mathcal{D}_{\gamma_2})$ via

$$\widehat{\mathcal{A}}_\varepsilon^\tau\left(\mathcal{D}_{\gamma_1}, \mathcal{D}_{\gamma_2}\right) := \widehat{\partial C}_0\left(\mathcal{D}_{\gamma_1}, \mathcal{D}_{\gamma_2}\right)\left(\widehat{C}_0\left(\mathcal{D}_{\gamma_2}\right) + \varepsilon \,\mathrm{id}\right)^{-1}.$$

[7] Alternatively, one can use uniform weights, i.e., $\alpha = \beta = \frac{1}{|\mathcal{I}_1|}\mathbf{1}$, $z = z' = \frac{1}{|\mathcal{I}_2|}\mathbf{1}$. See Footnote 2 for details.

M. Constructing Eigenfunctions from Finite-Dimensional Matrices

In this section, we construct eigenfunctions of the sparse approximation of Koopman operator using Gram matrices over sparse data. To that end, define

$$\Phi_X = [\phi(x_1), \ldots, \phi(x_d)], \quad \Phi_{X^+} = \left[\phi\left(x_1^+\right), \ldots, \phi(x_d^+)\right],$$

where $d = |\mathcal{D}_\gamma|$. Recall that the regularized sparse estimator of Koopman operator is given by $\widehat{\mathcal{K}}_\varepsilon = \left(\widehat{C}_{XX} + \varepsilon \,\mathrm{id}\right)^{-1} \widehat{C}_{XX^+}$. The sparse covariance and the cross-covariance operators in $\widehat{\mathcal{K}}_\varepsilon$ can be written as

$$\widehat{C}_{XX^+} = \Phi_X A_\alpha \Phi_{X^+}^\top, \quad \widehat{C}_{XX} = \Phi_X A_\beta \Phi_X^\top,$$
$$A_\alpha = \mathrm{diag}\,(\boldsymbol{\alpha}), \quad A_\beta = \mathrm{diag}\,(\boldsymbol{\beta}),$$

where α and β are as defined in Section 3. Define the Gram matrices

$$G_{XX} := \Phi_X^\top \Phi_X, \qquad G_{X^+X} := \Phi_{X^+}^\top \Phi_X, \tag{104}$$

so that $[G_{XX}]_{ij} = \kappa(x_i, x_j)$, and $[G_{X^+X}]_{ij} = \kappa(x_i^+, x_j^+)$, and both G_{XX} and G_{X^+X} are elements of $\mathbb{R}^{d\times d}$. Using this notation, rewrite $\widehat{\mathcal{K}}_\varepsilon$ as

$$\begin{aligned}
\widehat{\mathcal{K}}_\varepsilon &= \left(\widehat{C}_{XX} + \varepsilon \,\mathrm{id}\right)^{-1} \widehat{C}_{XX^+} \\
&= \left(\Phi_X A_\beta \Phi_X^\top + \varepsilon \,\mathrm{id}\right)^{-1} \Phi_X A_\alpha \Phi_{X^+}^\top \\
&= \Phi_X \left(A_\beta \Phi_X^\top \Phi_X + \varepsilon I\right)^{-1} A_\alpha \Phi_{X^+}^\top \\
&= \Phi_X \underbrace{\left(A_\beta G_{XX} + \varepsilon I\right)^{-1} A_\alpha}_{\Upsilon} \Phi_{X^+}^\top \\
&= \Phi_X \Upsilon \Phi_{X^+}^\top,
\end{aligned} \tag{105}$$

where $\Upsilon := (A_\beta G_{XX} + \varepsilon I)^{-1} A_\alpha$, and the third equality follows from the identity

$$(I + PQ)^{-1} P = P(I + QP)^{-1}. \tag{106}$$

From Klus et al. (2020c, Proposition 3.1), we get that an operator of the form $\widehat{\mathcal{K}}_\varepsilon = \Phi_X \Upsilon \Phi_{X^+}^\top$ has an eigenvalue λ with the corresponding eigenfunction

$$\varphi_\lambda(x) = (\Phi_X \boldsymbol{v})(x) = \sum_{i=1}^{d} v_i \kappa(x_i, x) \tag{107}$$

if and only if $\boldsymbol{v} = [v_1, \ldots, v_d]^\top$ is a right eigenvector of ΥG_{X^+X} associated with the same eigenvalue, i.e., $\Upsilon G_{X^+X} \boldsymbol{v} = \lambda \boldsymbol{v}$. Therefore, eigenfunctions of $\widehat{\mathcal{K}}_\varepsilon$ can be obtained by solving the finite-dimensional eigenvalue problem $\Upsilon G_{X^+X} \boldsymbol{v} = \lambda \boldsymbol{v}$ based upon finite-dimensional Gram matrices G_{XX} and G_{X^+X}.

Along the same lines as above, we next construct the generator using finite dimensional matrices. Define

$$\Phi_X = [\phi(x_1), \ldots, \phi(x_d)], \quad \mathrm{d}^{(2)}\Phi_X = \left[\mathrm{d}^{(2)}\phi\left(x_1\right), \ldots, \mathrm{d}^{(2)}\phi(x_d)\right],$$

where $d = |\mathcal{D}_\gamma|$. Recall that the regularized sparse approximation of PF generator is given by $\widehat{\mathcal{A}}_\varepsilon = \widehat{\partial C}_0 \left(\widehat{C}_0 + \varepsilon \,\mathrm{id}\right)^{-1}$. Per (34), we have

$$\widehat{\partial C}_0 = \mathrm{d}^{(2)}\Phi_X A_z \Phi_X^\top, \quad \widehat{C}_0 = \Phi_X A_{z'} \Phi_X^\top,$$
$$A_z = \mathrm{diag}\,(\boldsymbol{z}), \quad A_{z'} = \mathrm{diag}\,(\boldsymbol{z'}).$$

Define finite dimensional matrices

$$G_{10} = \mathrm{d}^{(2)}\Phi_X^\top \Phi_X, G_{00} = \Phi_X^\top \Phi_X, \tag{108}$$

where, due to the partial derivative reproducing property, entries of matrix G_{10} are given by

$$\begin{aligned}
[G_{10}]_{pq} &= \left\langle \mathrm{d}^{(2)}\phi(x_p), \phi(x_q) \right\rangle_{\mathcal{H}} \\
&= \left\langle \sum_{i=1}^{n} b_i(x_p) D^{e_i}\phi(x_p) + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}(x) D^{e_i+e_j}\phi(x_p), \phi(x_q) \right\rangle_{\mathcal{H}} \\
&= \sum_{i=1}^{n} b_i(x_p) \left\langle D^{e_i}\phi(x_p), \phi(x_q) \right\rangle_{\mathcal{H}} + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}(x_p) \left\langle D^{e_i+e_j}\phi(x_p), \phi(x_q) \right\rangle_{\mathcal{H}} \\
&= \sum_{i=1}^{n} b_i(x_p) D^{(e_i,0)}\kappa(x_p, x_q) + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}(x_p) D^{(e_i+e_j,0)}\kappa(x_p, x_q).
\end{aligned} \tag{109}$$

Using this notation and following the same logic as the previous case, the eigenfunction of $\widehat{\mathcal{A}}_\varepsilon$ has an eigenvalue λ with the corresponding eigenfunction

$$\varphi_\lambda(x) = (\Phi_X G_{00}^{-1}\boldsymbol{v})(x) \tag{110}$$

if and only if $\boldsymbol{v}$ is a right eigenvector of $A_z G_{10}\left(A_{z'} G_{00} + \varepsilon I\right)^{-1}$ associated with the same eigenvalue.

N. Partial Derivatives of the Gaussian Kernel

Consider the Gaussian kernel $\kappa(x, y) = \exp\left(-\|x-y\|^2/(2\sigma^2)\right)$ with $\sigma > 0$. The first-order partial derivative of $\kappa(x, y)$ w.r.t x is given by

$$D^{(e_i,0)}\kappa(x, y) = -\frac{1}{\sigma^2}(x_i - y_i)\kappa(x, y).$$

Applying the partial derivative reproducing property, if $i = j$, then we have

$$\begin{aligned}
D^{(e_i,e_i)}\kappa(x, y) &= \left\langle (D^{e_i}\kappa)_x, (D^{e_i}\kappa)_y \right\rangle \\
&= \left\langle \lim_{h\to 0} \frac{\kappa(x + he_i, \cdot) - \kappa(x, \cdot)}{h}, (D^{e_i}\kappa)_y \right\rangle \\
&\overset{(a)}{=} \lim_{h\to 0} \left\langle \frac{\kappa(x + he_i, \cdot) - \kappa(x, \cdot)}{h}, D^{(0,e_i)}\kappa(\cdot, y) \right\rangle \\
&\overset{(b)}{=} \lim_{h\to 0} \frac{D^{(0,e_i)}\kappa(x + he_i, y) - D^{(0,e_i)}\kappa(x, y)}{h} \\
&= \lim_{h\to 0} \frac{\frac{1}{\sigma^2}(x_i + h - y_i)\kappa(x + he_i, y) - \frac{1}{\sigma^2}(x_i - y_i)\kappa(x, y)}{h} \\
&= \lim_{h\to 0} \left(\frac{\frac{1}{\sigma^2} h\kappa(x + he_i, y)}{h} + \frac{\frac{1}{\sigma^2}\left((x_i - y_i)\kappa(x + he_i, y) - (x_i - y_i)\kappa(x, y)\right)}{h} \right) \\
&= \frac{1}{\sigma^2}\lim_{h\to 0}\kappa(x + he_i, y) + \frac{1}{\sigma^2}(x_i - y_i)\lim_{h\to 0}\frac{[\kappa(x + he_i, y) - \kappa(x, y)]}{h} \\
&= \frac{1}{\sigma^2}\lim_{h\to 0}\kappa(x + he_i, y) + \frac{1}{\sigma^2}(x_i - y_i)D^{(e_i,0)}\kappa(x, y) \\
&= \left(\frac{1}{\sigma^2} - \frac{1}{\sigma^4}(x_i - y_i)^2\right)\kappa(x, y),
\end{aligned} \tag{111}$$

where in line (a), we use the continuity of inner product to exchange limit with inner product and (b) follows from the partial derivative reproducing property. When $i \neq j$,

$$\begin{aligned}
D^{(e_i,e_j)}\kappa(x,y) &= \left\langle (D^{e_i}\kappa)_x, (D^{e_j}\kappa)_y \right\rangle \\
&= \left\langle \lim_{h\to 0} \frac{\kappa(x+he_i,\cdot)-\kappa(x,\cdot)}{h}, (D^{e_j}\kappa)_y \right\rangle \\
&\overset{(a)}{=} \lim_{h\to 0} \left\langle \frac{\kappa(x+he_i,\cdot)-\kappa(x,\cdot)}{h}, D^{(0,e_j)}\kappa(\cdot,y) \right\rangle \\
&\overset{(b)}{=} \lim_{h\to 0} \frac{D^{(0,e_j)}\kappa(x+he_i,y)-D^{(0,e_j)}\kappa(x,y)}{h} \\
&= \lim_{h\to 0} \frac{\frac{1}{\sigma^2}(x_j-y_j)\kappa(x+he_i,y)-\frac{1}{\sigma^2}(x_j-y_j)\kappa(x,y)}{h} \\
&= \lim_{h\to 0} \frac{1}{\sigma^2}(x_j-y_j)\frac{\kappa(x+he_i,y)-\kappa(x,y)}{h} \\
&= \frac{1}{\sigma^2}(x_j-y_j)\lim_{h\to 0} \frac{\kappa(x+he_i,y)-\kappa(x,y)}{h} \\
&= \frac{1}{\sigma^2}(x_j-y_j)D^{(e_i,0)}\kappa(x,y) \\
&= -\frac{1}{\sigma^4}\left(x_i-y_i\right)\left(x_j-y_j\right)\kappa(x,y).
\end{aligned} \tag{112}$$

Likewise, the second-order derivatives are given by

$$D^{(e_i+e_j,0)}\kappa\left(x,y\right) = \begin{cases} \left(\frac{1}{\sigma^4}\left(x_i-y_i\right)^2-\frac{1}{\sigma^2}\right)\kappa\left(x,y\right) & i=j, \\ \frac{1}{\sigma^4}\left(x_i-y_i\right)\left(x_j-y_j\right)\kappa\left(x,y\right) & i\neq j, \end{cases}$$

$$D^{(e_i+e_j,e_p)}\kappa(x,y) = \begin{cases} \frac{1}{\sigma^4}\left(\frac{1}{\sigma^2}\left(x_p-y_p\right)^2-3\right)\left(x_p-y_p\right)\kappa(x,y) & i=j=p, \\ \frac{1}{\sigma^4}\left(\frac{1}{\sigma^2}\left(x_p-y_p\right)^2-1\right)\left(x_q-y_q\right)\kappa(x,y) & i\neq j, p\neq q, q\in\{i,j\} \\ \frac{1}{\sigma^4}\left(\frac{1}{\sigma^2}\left(x_i-y_i\right)^2-1\right)\left(x_p-y_p\right)\kappa(x,y) & i=j\neq p, \end{cases}$$

$$D^{(e_i+e_j,e_p+e_q)}\kappa(x,y) = \begin{cases} \left(\frac{1}{\sigma^6}\left(\frac{(x_p-y_p)^2}{\sigma^2}-6\right)(x_p-y_p)^2+\frac{3}{\sigma^4}\right)\kappa(x,y) & i=j=p=q, \\ \left(\frac{1}{\sigma^6}\left(\frac{(x_p-y_p)^2}{\sigma^2}-3\right)(x_p-y_p)(x_i-y_i)\right)\kappa(x,y) & i\neq j, p=q, \\ \left(\frac{1}{\sigma^6}\left(\frac{(x_p-y_p)^2}{\sigma^2}-1\right)(x_i-y_i)^2+\frac{1}{\sigma^4}-\frac{1}{\sigma^6}(x_p-y_p)^2\right)\kappa\left(x,y\right), & i=j\neq p=q, \\ \left(\frac{1}{\sigma^6}\left(\frac{(x_i-y_i)^2}{\sigma^2}-3\right)(x_p-y_p)(x_i-y_i)\right)\kappa(x,y), & i=j, p\neq q, \\ \left(\frac{1}{\sigma^6}\left(\frac{(x_p-y_p)^2}{\sigma^2}-1\right)(x_i-y_i)^2+\frac{1}{\sigma^4}-\frac{1}{\sigma^6}(x_p-y_p)^2\right)\kappa\left(x,y\right) & i\neq j, p\neq q. \end{cases}$$