Propagating Uncertainty Through System Dynamics in Reproducing Kernel Hilbert Space

Boya Hou[a,*], Amarsagar Reddy Ramapuram Matavalam[b], Subhonmesh Bose[a], Umesh Vaidya[c]

[a] *Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, 61801, IL, United Stated*
[b] *School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, 85287, AZ, United Stated*
[c] *Department of Mechanical Engineering, Clemson University, Clemson, 29634, SC, United Stated*

Abstract

We present a data-driven approach to propagate uncertainty in initial conditions through the dynamics of an unknown system in a reproducing kernel Hilbert space (RKHS). The uncertainty in initial conditions is represented through its kernel mean embedding (KME) in the RKHS. For a discrete-time Markovian dynamical system, we utilize the conditional mean embedding (CME) operator to encode the underlying dynamics. Learning in RKHS often incurs prohibitive data storage requirements. To circumvent said limitation, we propose an algorithm to propagate uncertainty via a learned *sparse* CME operator, and provide theoretical guarantees on the approximation error for the embedded distribution with time. We empirically study our algorithm over illustrative dynamical systems and power systems.

Keywords: dynamical system, uncertainty propagation, RKHS, sparse approximation

*Corresponding author

Email addresses: boyahou2@illinois.edu (Boya Hou), amar.sagar@asu.edu (Amarsagar Reddy Ramapuram Matavalam), boses@illinois.edu (Subhonmesh Bose), uvaidya@clemson.edu (Umesh Vaidya)

1. Introduction

A dynamical system describes how states of a system evolve over time. When the initial point of the system is uncertain, one can represent this uncertainty through a probability distribution or its support. Uncertainty propagation entails tracking the evolution of the probability distribution or the support of the states over time under the action of the system dynamics. A study of this distribution of states over time naturally has important applications in the analysis of stability and robustness of the system. For example, perception errors of LiDAR sensors in autonomous vehicles affect accuracy of state estimation. Said estimation errors in initial points will lead to deviations of actual trajectories from the planned trajectory. In turn, such deviations can impact the safety of the vehicle. To avoid such possibilities, one must consider and quantify the impact of erroneous sensor readings by studying the range of potential trajectories to design a sound motion plan. Another potential application for uncertainty propagation is transient stability analysis in power system operations. When a fault occurs in a power system, one must clear that fault to isolate its effects before the on-fault trajectory of the system leaves the region of attraction of a stable equilibrium point of the post-fault dynamics. Such analysis is often done offline and is used to set settings for relays/circuit-breakers to clear the fault. The initial point for the on-fault trajectory used for the study may not coincide with the point where the fault occurs in practice. One must account for these deviations in designing relay settings.

Methods for uncertainty propagation often require an explicit model of the system dynamics, often represented as ordinary differential equations (ODEs) or differential-algebraic equations (DAEs) (see Choi et al. [1], Chen and Domínguez-García [2], Pico et al. [3], Jiang and Domínguez-García [4]). For complex systems, a succinct mathematical description of the dynamical system can be difficult to obtain. Monte Carlo-based methods in Halton [5], Hanson [6], Helton [7] circumvent that difficulty and only rely on a simulator that can generate trajectories, given initial points. These methods have been known to be data-intensive and often computationally prohibitive, per Mezic and Runolfsson [8], Xu et al. [9], Matavalam et al. [10]. A viable alternative to Monte Carlo simulations is *polynomial chaos* advocated in Ghanem and Red-Horse [11], Ghanem and Spanos [12], Xu et al. [9]. While it is computationally more efficient than Monte Carlo methods, its major drawback is that the whole computation needs to be restarted every time one needs to study the

propagation of different probability distributions for the uncertainty in the initial condition. To circumvent this limitation, one must suitably learn a representation of the system dynamics. Recently, Matavalam et al. [10, 13] proposed a mechanism to propagate the moments of the distributions of the states through time using a data-driven approximation to the Koopman operator.

Transfer operators such as the Koopman and the Perron-Frobenius operators (see Lasota and Mackey [14]) essentially describe how distributions and functions of states of a system evolve through time. These infinite-dimensional yet linear operators have a long history in the study of dynamical systems (see Mezić and Banaszuk [15], Mezić [16], Mauroy and Mezić [17]). They have gained in popularity recently as they naturally lend themselves to approximations from data collected from trajectories of the system. Such data-driven analysis requires one to study the action of these operators with suitable function spaces. The extended dynamic mode decomposition (EDMD) method in Williams et al. [18], Klus et al. [19] is a data-driven algorithm that seeks to learn these operators via their actions on parameterized function spaces. See Korda and Mezić [20] for asymptotic convergence results for the EDMD method. Other approaches such as (deep) neural networks (Li et al. [21], Yeung et al. [22], Takeishi et al. [23], Lusch et al. [24], Wehmeyer and Noé [25], Otto and Rowley [26]) have also been utilized to parameterize these function spaces. In Matavalam et al. [10], the authors pre-select a basis for this function space, use this space to learn an approximate Koopman operator, and finally leverage the learned approximation to propagate moments of the distributions that describe the uncertainties in the system states through time. Given the challenges of selecting a pre-defined basis for the function spaces, we take the non-parametric route advocated in Williams et al. [27], Klus et al. [28, 29], Hou et al. [30], Kostic et al. [31], Meanti et al. [32] to learn the transfer operators from data in the reproducing kernel Hilbert space (RKHS). Along a similar line, we propose a *non-parametric* data-driven approach to represent and propagate uncertainty in initial conditions. We rely on the actions of these operators in RKHS. Specifically, we embed the probability distribution of initial uncertainty sets into an RKHS. Given sampled snapshot pairs of initial points and their next state propagated through the dynamics, we build a model of the underlying dynamical system as the so-called *conditional mean embedding (CME)* operator that acts on an RKHS. CMEs capture the transition dynamics without resorting to explicit mathematical representation of the system dynamics such as those via ordi-

nary/stochastic differential or differential-algebraic equations. RKHS theory is mature and it has found widespread applications in statistical learning theory (see Steinwart and Christmann [33], Berlinet and Thomas-Agnan [34]); the CME framework that builds on properties of the RKHS has also found several applications (see Song et al. [35], Muandet et al. [36], Hou et al. [37] that include our own prior works). In a nutshell, this representation is fully data-oriented and reduces computations of high-dimensional integrations required for uncertainty propagation to simple inner products, calculations of which are independent of the dimension of the state. Moreover, convergence guarantees on data-driven estimates of the various embeddings are available, see Song et al. [35], Hou et al. [38]. Representations in the RKHS, however, become prohibitive as the dataset grows without bounds. In order to improve the scalability of the kernel method, Engel et al. [39], Kivinen et al. [40], Wu et al. [41], Rahimi and Recht [42], Koppel et al. [43], Chatalic et al. [44] propose a variety of sparsification procedures to reduce redundancy in the dataset. In this paper, we control the model complexity via the coherence condition introduced by Richard et al. [45] and build a sparse model of the underlying dynamics via sparse CME based on our prior work in Hou et al. [30].

Our key contributions are as follows: (1) We present a non-parametric algorithm to propagate uncertainty in initial conditions through unknown nonlinear dynamics. In particular, we control the complexity of the learned model and use such a sparse model to propagate the embedded probability distribution of initial states. (2) We bound the error in propagating the embedded uncertainty. Such non-asymptotic error analysis indicates that the control of model complexity comes at the price of accuracy, and cannot be avoided simply with more samples.

Closest in spirit to our algorithm is the paper by Zhu et al. [46]. One important aspect in which we differ from Zhu et al. [46] is that we learn a *sparse* CME operator for uncertainty propagation. The process of sparsification is essential since computations with kernels become increasingly burdensome with the number of samples used for learning. Another aspect that distinguishes our work from Zhu et al. [46] is that we provide an explicit non-asymptotic error analysis. Specifically, leveraging the theoretical results in Hou et al. [30], we provide theoretical guarantees on the approximation errors for propagated embedded probability distributions. Our result illustrates the relationship between the number of samples and the level of sparsification on the resulting error performance.

The rest of the paper is organized as follows. Section 2.1,2.2 and 3.1 serve as prerequisites for learning dynamical systems in RKHS. The main algorithm is presented in Section 3.2, followed by theoretical analysis. Our approach proceeds in two stages: given sampled snapshot pairs, we start by constructing a sparse CME estimator offline. After receiving initial samples, we compute an embedding of the initial state distribution and then evolve that through the learned CME operator. By doing so, uncertainty propagation reduces to computationally efficient matrix multiplications. Finally, we demonstrate the efficacy of our algorithm for simple dynamical systems in Section 4 and specifically for example power systems in Section 5.

2. Preliminaries

2.1. RKHS and Conditional Mean Embedding

We first briefly review some definitions and properties regarding RKHS and embeddings of probability distributions. Consider a Euclidean space $\mathbb{X} \subseteq \mathbb{R}^n$. Let $\kappa : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ be a symmetric positive semi-definite kernel function with feature map $\phi(x) := \kappa(x, \cdot)$. κ defines an RKHS $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ as the closure of span $\{\phi(x) := \kappa(x, \cdot) : x \in \mathbb{X}\}$ with respect to the inner product $\langle \kappa(x, \cdot), \kappa(y, \cdot) \rangle_{\mathcal{H}} = \kappa(x, y)$. In particular, the $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ has the reproducing property, given by

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

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and consider a $\mathbb{X}$-valued random variable $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{X}, \Sigma, \mathbb{P}_X)$, where Σ is the Borel σ-algebra on $\mathbb{X}$ and $\mathbb{P}_X$ is a distribution on $\mathbb{X}$. One can then embed the marginal probability distribution $\mathbb{P}_X$ into $\mathcal{H}$. To be precise, if κ is $\Sigma \times \Sigma$-measurable, and $\mathbb{E}_X \left[\sqrt{\kappa(x, x)} \right] < \infty$, then there exists a *kernel mean embedding* (KME) $\mu : \mathbb{P}_X \mapsto \mu_{\mathbb{P}_X} \in \mathcal{H}$ such that

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

Throughout this paper, we suppose that κ is continuous and bounded as $\sup_{x \in \mathbb{X}} \kappa(x, x) \leq B_\kappa < \infty$ for some $B_\kappa \in \mathbb{R}_+$ so that the KME is well-defined as an element in $\mathcal{H}$, per Muandet et al. [36, Lemma 3.1].

Let $Y : (\Omega', \mathcal{F}', \mathbb{P}') \to (\mathbb{Y}, \Sigma_Y, \mathbb{P}_Y)$ be a $\mathbb{Y}$-valued random variable and $\mathcal{H}_1, \mathcal{H}_2$ be two RKHSs on $\mathbb{X}$ and $\mathbb{Y}$ with kernel functions κ_1,κ_2, respectively.

One can define a tensor product Hilbert space $\mathcal{H}_1 \otimes \mathcal{H}_2$ with kernel function

$$\kappa_{\otimes}\Big((x_1, y_1), (x_2, y_2)\Big) = \kappa_1(x_1, x_2)\ \kappa_2(y_1, y_2), \tag{3}$$

for all $x_1, x_2 \in \mathbb{X}$, $y_1, y_2 \in \mathbb{Y}$ and (joint) feature map

$$\varphi\left(x_i, y_i\right) := \phi_1\left(x_i\right) \otimes \phi_2\left(y_i\right) = \kappa_1\left(x_i, \cdot\right) \kappa_2\left(y_i, \cdot\right). \tag{4}$$

A joint distribution $\mathbb{P}_{XY}$ can be embedded into $\mathcal{H}_1 \otimes \mathcal{H}_2$ as

$$C_{XY} = \mathbb{E}_{XY}[\phi_1(X) \otimes \phi_2(Y)], \tag{5}$$

where C_{XY} is the (uncentered) *cross-covariance* operator. Alternatively, C_{XY} can also be viewed as a Hilbert-Schmidt (HS) operator $C_{XY} : \mathcal{H}_2 \to \mathcal{H}_1$ that satisfies

$$\mathbb{E}_{XY}[f(X)g(Y)] = \langle C_{XY} g, f\rangle_{\mathcal{H}}, \quad \forall f \in \mathcal{H}_1, g \in \mathcal{H}_2, \tag{6}$$

per Berlinet and Thomas-Agnan [34]. Likewise, one can embed the marginal distribution $\mathbb{P}_X$ into $\mathcal{H}_1 \otimes \mathcal{H}_1$ as

$$C_{XX} := \mathbb{E}_X[\phi_1(X) \otimes \phi_1(X)]. \tag{7}$$

With a slight abuse of notion, we denote the joint feature map of $\mathcal{H}_1 \otimes \mathcal{H}_1$ as $\varphi\left(x_i, x_i\right) := \phi_1\left(x_i\right) \otimes \phi_1\left(x_i\right) = \kappa_1\left(x_i, \cdot\right) \kappa_1\left(x_i, \cdot\right)$.

Yet embeddings of marginal distribution cannot capture the dependency between random variables. To this end, the conditional mean embedding embeds conditional probability distributions into RKHS and encodes how distribution over one random variable relates to another. Specifically, let $\mathbb{P}_{Y|x}$ denote the conditional distribution of the random variable Y given $X = x \in \mathbb{X}$. The embedding of $\mathbb{P}_{Y|x}$ into $\mathcal{H}_2$ is the Bochner conditional expectation.

$$\mu_{\mathbb{P}_{Y|x}} := \mathbb{E}_{Y|x}[\phi_2(Y)|X = x] \qquad \forall x \in \mathbb{X}. \tag{8}$$

Under the prevalent definition given by Song et al. [35], the conditional mean embedding operator $\mathcal{U} : \mathcal{H}_1 \to \mathcal{H}_2$, is a linear operator that satisfies

$$\mu_{\mathbb{P}_{Y|x}} = \mathcal{U}\phi_1(x). \tag{9}$$

In addition, if $\mathbb{E}_{Y|x}[f(Y)|X = x] \in \mathcal{H}_2$ for all $f \in \mathcal{H}_2$ and $x \in \mathbb{X}$, then we have

$$\mathcal{U} = C_{YX} C_{XX}^{\dagger}. \tag{10}$$

For technical reasons concerning inverting a linear operator, we consider the regularized version, defined as

$$\mathcal{U}_{\varepsilon} = C_{YX} \left(C_{XX} + \varepsilon \, \mathrm{Id} \right)^{-1}, \tag{11}$$

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

We remark that the assumption $\mathbb{E}_{Y|x}[f(Y)|X = x] \in \mathcal{H}_2$ for all $f \in \mathcal{H}_2$ requires the RKHS to be closed with respect to the evolution of functions through the system dynamics. Such closedness assumption is not new to applications utilizing the CME operator (see Song et al. [35, 47], Klus et al. [28], Hou et al. [38]) and Koopman operator-based analysis (Yeung et al. [22], Nandanoori et al. [48]). According to Song et al. [35], this assumption holds for finite domains with a characteristic kernel; see Muandet et al. [36] for a definition of a characteristic kernel. Yet as noted by Park and Muandet [49], Klebanov et al. [50], this assumption can be restrictive in a more general setting. In light of recent developments in Li et al. [51], when the assumption does not hold, $\mathcal{U}_{\varepsilon}$ defined in (11) can be viewed as an approximation of the true CME operator $\mathcal{U}$ for which the approximation is always off by a term that encodes how far the operator is from the hypothesis space.

2.2. CMEs for Discrete-Time Dynamical Systems

Let $\mathbb{N}$ be the set of nonnegative integers and $\{X_t\}_{t \in \mathbb{T}}$ be a discrete time dynamical system described by the recursion

$$x_{t+1} = F(x_t, \omega_t), \quad t \in \mathbb{T}, \tag{12}$$

where the mapping $F : \mathbb{X} \times \mathbb{W} \to \mathbb{X}$ is diffeomorphism and ω_t are i.i.d. random variable with values in $\mathbb{W}$ that are independent of X_0. Such dynamics can also be specified by the transition kernel density p as [1]

$$\mathbb{P}\left(X_{t+1} \in \mathbb{A} | X_t = x\right) = \int_{\mathbb{A}} p(y|x) \mathrm{d}y, \tag{13}$$

[1] For a deterministic dynamical system of the form $x_{t+1} = F(x_t)$, it can be described under the more general framework (13) with $\mathbb{P}(\cdot|x)$ being the Dirac delta measure $\delta_{F(x)}$ supported on the point $x^+ = F(x)$, per Dellnitz and Junge [52].

for measurable $\mathbb{A} \subseteq \mathbb{X}$. If f is a probability density over $\mathbb{X}$, then the Perron–Frobenius (PF) operator $\mathcal{P}$ propagates f as

$$(\mathcal{P}f)(y) = \int p(y|x)f(x)\mathrm{d}x. \tag{14}$$

Recall from Section 2.1 that CME encodes how the probability distribution over one random variable relates to another. If the random variables correspond to successive states of a discrete-time dynamical system, the CME operator naturally captures the state transition dynamics. In this regard, one can identify the PF operator as the CME operator, since $\mathcal{U} : \mu_{\mathbb{P}_X} \mapsto \mu_{\mathbb{P}_{X^+}}$ satisfies

$$\mu_{X^+} = \mathcal{U}\mu_X, \tag{15}$$

per Song et al. [35] and Hou et al. [30], where x^+ is the system state at the next time-step starting from x. In other words, the embedded PF operator $\mathcal{P}$ is the CME operator that propagates the embedded distribution of states through the system dynamics as Figure 1 illustrates.

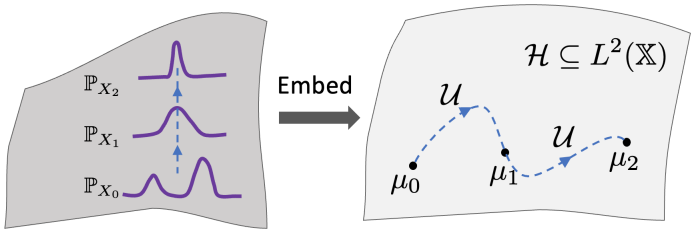


Figure 1: The CME operator $\mathcal{U}$ propagates the embedded distribution of states μ through system dynamics.

3. Propagating Uncertainty In Initial Conditions

We now present our method to propagate uncertainty in initial conditions through system dynamics, where the uncertainty is described by a probability distribution on $\mathbb{X}$. Our approach proceeds in two stages. First, we learn a sparse representation of the system dynamics via the CME operator. Then, after receiving samples of uncertain initial points, we define an empirical estimate of the distribution using KME and then evolve it through the learned sparse CME operator.

3.1. Sparse Learning of Dynamical System in RKHS

Consider an RKHS $\mathcal{H}$ associated with kernel $\kappa : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ and feature map $\phi : \mathbb{X} \to \mathcal{H}$. Given M snapshot pairs $\mathcal{D} := \left\{ \left(x_1, x_1^+\right), \ldots, \left(x_M, x_M^+\right) \right\}$ sampled from $\mathbb{P}_{XX^+}$, the empirical estimates of C_{XX} and C_{XX^+} are

$$\widetilde{C}_{XX} = \frac{1}{M} \sum_{i=1}^{M} \varphi\left(x_i, x_i\right), \quad \widetilde{C}_{XX^+} = \frac{1}{M} \sum_{i=1}^{M} \varphi\left(x_i, x_i^+\right), \tag{16}$$

The regularized empirical estimate of the CME operator can be defined as

$$\widetilde{\mathcal{U}}_\varepsilon := \widetilde{C}_{X^+X} \left(\widetilde{C}_{XX} + \varepsilon \, \mathrm{Id} \right)^{-1}. \tag{17}$$

The sample complexity of (16) with independently and identically distributed data is well understood (see Muandet et al. [36]). While the estimation accuracy improves with more samples, the increase in computational complexity makes learning in RKHS computationally burdensome when scales to large data, as mentioned by Engel et al. [39, 53], Kivinen et al. [40], Hou et al. [30]. To efficiently estimate the CME operator, we leverage the notion of *coherency* proposed by Richard et al. [45] to identify a subset of $\mathcal{D}$ based on which a sparse CME operator is constructed. Informally, one throws away those points in $\mathcal{D}$ that are deemed "look-alike" with respect to κ. To be precise, we prune $\mathcal{D}$ to construct a sparse dictionary $\widehat{\mathcal{D}}$ such that

$$\left|\kappa(x_i^\star, x_j^\star)\right| \leq \sqrt{\gamma \kappa(x_i^\star, x_i^\star) \kappa(x_j^\star, x_j^\star)}, \tag{18}$$

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 $\widehat{\mathcal{D}}$. The sparse dictionary $\widehat{\mathcal{D}}$ can be constructed as follows. Consider Gram matrices computed using all elements in $\mathcal{D}$ with kernel function κ. Since (18) consists of two conditions, i.e., one for (x_i, x_j) and one for (x_i^+, x_j^+), the process involves two Gram matrices K and K^+ whose elements are given be $K_{ij} = \kappa(x_i, x_j)$ and $K_{ij}^+ = \kappa(x_i^+, x_j^+)$. For two indices s, t with $s < t$, if (x_s, x_s^+) and (x_t, x_t^+) violate (18), we keep (x_s, x_s^+) in $\mathcal{D}$, but throw away (x_t, x_t^+). At the same time, we delete the row/column associated with x_t from both K and the row/column associated with x_t^+ from K^+. We then repeat such a process until all elements in the Gram matrices satisfy (18) to obtain $\widehat{\mathcal{D}}$. To compute the sparse cross-covariance and covariance estimators $\widehat{C}_{X^+X}$

and $\widehat{C}_{XX}$ based on $\widehat{\mathcal{D}}$, let $\mathcal{I}$ be the indices among $1, \ldots, M$ for which (x_i, x_i^+) are in $\widehat{\mathcal{D}}$. Then, we define the sparse estimators as

$$\widehat{C}_{X^+X} = \sum_{i \in \mathcal{I}} \alpha_i \varphi\left(x_i^+, x_i\right), \quad \widehat{C}_{XX} = \sum_{i \in \mathcal{I}} \beta_i \varphi\left(x_i, x_i\right), \tag{19}$$

where α (and similarly, β) solves

$$\min_{\overline{\alpha}} \left\| \frac{1}{m} \sum_{i=1}^{M} \varphi\left(x_i^+, x_i\right) - \sum_{i \in \mathcal{I}} \overline{\alpha}_i \varphi\left(x_i^+, x_i\right) \right\|_{\mathcal{H} \otimes \mathcal{H}}^2. \tag{20}$$

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

$$G_{i,j} = \kappa\left(x_i^+, x_j^+\right) \kappa\left(x_i, x_j\right), \tag{21}$$

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

$$[g]_j = \frac{1}{M} \sum_{i=1}^{M} G_{ij}, \quad j \in \mathcal{I}. \tag{22}$$

Using $\widehat{C}_{XX}$ and $\widehat{C}_{X^+X}$, we compute the *regularized sparse CME estimator* by

$$\widehat{\mathcal{U}}_\varepsilon := \widehat{C}_{X^+X} \left(\widehat{C}_{XX} + \varepsilon \, \mathrm{Id}\right)^{-1}. \tag{23}$$

3.2. Uncertainty Propagation via CME

Having learned a sparse model encoded in $\widehat{\mathcal{U}}_\varepsilon$, we next apply it to propagate uncertainty in initial conditions. Given initial samples $\mathcal{Z} := (z_i)_{i=1}^N$ where $z \in \mathbb{X}$, the embedded distribution of initial state is computed as

$$\widehat{\mu}_0 = \sum_{i=1}^{N} w_0(i) \, \kappa\left(z_i, \cdot\right), \quad w_0(i) = \frac{1}{N}, \quad \forall i = 1, \cdots, N. \tag{24}$$

Let $d = |\widehat{\mathcal{D}}|$, $A_{X^+X} := \mathrm{diag}\left(\boldsymbol{\alpha}\right), A_{XX} := \mathrm{diag}\left(\boldsymbol{\beta}\right)$ and define feature matrices

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

$\widehat{C}_{X^+X}, \widehat{C}_{XX}$ in (19) can then be rewritten as

$$\widehat{C}_{X^+X} = \Phi_{X^+} A_{X^+X} \Phi_X^\top, \quad \widehat{C}_{XX} = \Phi_X A_{XX} \Phi_X^\top.$$

Using the above representations, the one-step propagation of embedded state distribution can be computed as

$$\begin{aligned}
\widehat{\mu}_1 &:= \widehat{\mathcal{U}}_\varepsilon \, \widehat{\mu}_0 \\
&= \widehat{C}_{X^+X} \left(\widehat{C}_{XX} + \varepsilon \, \mathrm{Id} \right)^{-1} \widehat{\mu}_0 \\
&= \Phi_{X^+} A_{X^+X} \Phi_X^\top \left(\Phi_X A_{XX} \Phi_X^\top + \varepsilon \, \mathrm{Id} \right)^{-1} \widehat{\mu}_0 \\
&\overset{(a)}{=} \Phi_{X^+} \underbrace{A_{X^+X} \left(G_{XX} A_{XX} + \varepsilon \, \mathrm{Id} \right)^{-1} \Phi_X^\top \widehat{\mu}_0}_{:= w_1}, \\
&= \Phi_{X^+} w_1,
\end{aligned}$$

where $G_{XX} = \Phi_X^\top \Phi_X$ is the Gram matrix, line (a) follows from the identity $(I + PQ)^{-1} P = P(I + QP)^{-1}$, and the d-dimensional column vector w_1 is

$$w_1 := A_{X^+X} \left(G_{XX} A_{XX} + \varepsilon \, \mathrm{Id} \right)^{-1} \Phi_X^\top \widehat{\mu}_0.$$

Hence, we have

$$\widehat{\mu}_1 = \Phi_{X^+} w_1 = \sum_{i \in \mathcal{I}} w_1 (i) \, \kappa \left(x_i^+, \cdot \right).$$

Likewise, the embedded distribution of states at $t \geq 2$ can be computed as

$$\begin{aligned}
\widehat{\mu}_t &:= \widehat{\mathcal{U}}_\varepsilon \, \widehat{\mu}_{t-1} \\
&= \widehat{C}_{X^+X} \left(\widehat{C}_{XX} + \varepsilon \, \mathrm{Id} \right)^{-1} \widehat{\mu}_{t-1} \\
&= \Phi_{X^+} A_{X^+X} \Phi_X^\top \left(\Phi_X A_{XX} \Phi_X^\top + \varepsilon \, \mathrm{Id} \right)^{-1} \widehat{\mu}_{t-1} \\
&= \Phi_{X^+} A_{X^+X} \left(G_{XX} A_{XX} + \varepsilon \, \mathrm{Id} \right)^{-1} \Phi_X^\top \widehat{\mu}_{t-1} \\
&= \sum_{i \in \mathcal{I}} w_t (i) \, \kappa \left(x_i^+, \cdot \right),
\end{aligned}$$

where the vector w_t of coefficients is given by

$$w_t := A_{X^+X} \left(G_{XX} A_{XX} + \varepsilon \, \mathrm{Id} \right)^{-1} \Phi_X^\top \widehat{\mu}_{t-1}. \tag{25}$$

To summarize, the empirical estimate of embedded state distribution at time t can be constructed as a linear combination of $\{\kappa(x_i^+, \cdot)\}_{i\in\mathcal{I}}$, given by

$$\widehat{\mu}_t := \left(\widehat{\mathcal{U}}_\varepsilon\right)^t \widehat{\mu}_0 = \sum_{i\in\mathcal{I}} w_t(i)\,\kappa\left(x_i^+, \cdot\right), \quad t \in \mathbb{N}\setminus\{0\}, \tag{26}$$

where the evolution of the weight vector $w_t, t \in \mathbb{N}$ captures the evolution of the uncertainty in initial conditions through system dynamics.

3.2.1. Moment Propagation

With an estimation of embedded state distribution, one can also compute the a-th order moment of state distribution by taking an inner product with μ_t. To illustrate, consider a 1-dimensional random variable X. Suppose $x^a \in \mathcal{H}$ [2], then, we have

$$m_t^a := \mathbb{E}\left[X_t^a\right] = \langle X^a, \mu_t\rangle_{\mathcal{H}}. \tag{27}$$

Utilizing $\widehat{\mu}_t$, one can approximate m_t^a as

$$\begin{aligned}\widehat{m}_t^a = \langle X^a, \widehat{\mu}_t\rangle_{\mathcal{H}} &= \langle X^a, \sum_{i\in\mathcal{I}} w_t(i)\,\kappa\left(x_i^+, \cdot\right)\rangle_{\mathcal{H}} \\ &= \sum_{i\in\mathcal{I}} w_t(i)\,\langle X^a, \kappa\left(x_i^+, \cdot\right)\rangle_{\mathcal{H}} \\ &= \sum_{i\in\mathcal{I}} w_t(i)\left(x_i^+\right)^a.\end{aligned} \tag{28}$$

We summarize the algorithm in Algorithm 1.

3.3. Theoretical Analysis

We now present an upper bound on the estimation error of embedded state distribution for $t \in \mathbb{N}$. We use the notation $\|\cdot\|$ to denote operator norm, and

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

[2] When this is not satisfied, the derivation can be viewed as a formal approximation technique.

Algorithm 1 Uncertainty propagation in RKHS

Require: Kernel function κ; Snapshot pairs $\mathcal{D}$; Coherence parameter γ; Realizations of initial uncertainty set $\mathcal{Z}$.

1: Prune $\mathcal{D}$ to get γ-coherent dictionary $\widehat{\mathcal{D}}$ that satisfies (18)
2: Solve for coefficients $\boldsymbol{\alpha}, \boldsymbol{\beta}$ according to (19),(20).
3: At time $t = 0$, assign uniform weight to w_0 as (24).
4: **for** $t = 1, \cdots$ **do**
5: Find the coefficient vector w_t using (25)
6: Compute $\widehat{m}_t^a$ via (28)
7: **end for**

Theorem 1. *Assume that kernel κ is continuous and bounded as $\sup_{x\in\mathbb{X}} \kappa(x,x) \leq B_\kappa < \infty$ for some $B_\kappa \in \mathbb{R}_+$. Let $\mu_t := \mathcal{U}_\varepsilon^t \mu_0$ be the embedding of state distribution at time t. Given datasets $\mathcal{D}$ and $\mathcal{Z}$ that consist of M and N i.i.d. samples drawn according to $\mathbb{P}_{XX^+}$ and $\mathbb{P}_{X_0}$ respectively, μ_t and $\widehat{\mu}_t$ satisfies*

$$\|\mu_t - \widehat{\mu}_t\|_{\mathcal{H}} \leq B_\kappa^t \left[\frac{1}{\sqrt{N}} \Xi\left(\frac{\delta}{2}\right) \mathcal{O}\left(\varepsilon^{-t}\right) + t\psi\left(M, \gamma; \frac{\delta}{4}\right) \mathcal{O}\left(\varepsilon^{-(t+1)}\right) \right] \tag{29}$$

with probability at least $1-\delta$, for $t \in \mathbb{N}$ and $\delta \in [0,1]$, if $\widehat{C}_{XX}$ is positive semi-definite [3]*, and*

$$\psi(M, \gamma; \delta) := \frac{1}{\sqrt{M}} \Xi(\delta) + \left(1 - \frac{|\widehat{\mathcal{D}}|}{M}\right) \sqrt{1-\gamma^2}. \tag{30}$$

Our result suggests that the estimation error consists of two parts: the first term in (29) reflects the initial sampling error being propagated through the iteration which depends on $\sqrt{N}$, and the second term in (29) captures the error induced by sparse approximation of $\mathcal{U}_\varepsilon$. The former bears a resemblance to the global error of Euler's method in that the initial error increases with time exponentially. For the latter one, we leverage Hou et al. [30, Theorem 1] to obtain the approximation accuracy which depends on both the number of snapshot pairs M and the level of sparsity controlled by γ. In other words, the term $\sqrt{1-\gamma^2}$ encapsulates the price we pay for sparsity and the error

[3] Such an assumption is satisfied when coefficients $\beta \geq 0$.

due to sparsification grows linearly in t. Furthermore, our error estimate depends on the regularization parameter ε. A recent work by Li et al. [51] shows the potential to improve the dependency on ε that we plan to explore in future work. In addition, Theorem 1 implies that the estimation error of embedded state distribution is independent of the dimension of the state space–a hallmark property of kernel methods.

Theorem 1 is a "high probability" guarantee on the error in propagating uncertainty, much along the lines of Kostic et al. [31], Meanti et al. [32] derived for the Koopman operator. This result characterizes finite-sample performance and essentially has three parts to the error–one due to the variance from sampling initial points, second due to the variance of the point-pairs used to learn the Koopman operator, and third the bias from sparsification of the operator. Recall from Section 2.1 that our results are derived under a closedness condition that $\mathbb{E}_{X^+|x}[f(X^+)|X = x] \in \mathcal{H}$ for all $f \in \mathcal{H}$. A measure-theoretically sound alternate approach developed in Park and Muandet [49], Li et al. [51] avoids this requirement. With this definition, bounds on how well the operator itself can be learned can be developed as in Li et al. [51]. These bounds include an additional bias term from the inability to represent the target operator within the hypothesis space. Extending that analysis to a refined bound on learning the KME/CME for uncertainty propagation is left for future endeavors.

4. Illustrative Examples

We begin our numerical experiments with three simple dynamical systems and defer the power system examples to the next section. The performance of our method indeed depends upon the choice of the kernel function. We remark that even that choice can be optimized along the lines of Gretton et al. [54]. To capture all information of a probability distribution in its embedding, one requires the kernel to be characteristic, i.e., the mapping $\mathbb{P}_X \mapsto \mu_{\mathbb{P}_x}$ is injective so that $\mu_{\mathbb{P}_X}$ uniquely determines $\mathbb{P}_X$. See Muandet et al. [36] for details. In our experiments, we adopt Gaussian kernels that are known to be characteristic (see Fukumizu et al. [55]).

4.1. A 50-Dimensional Linear System

Consider a stochastic linear dynamical system,

$$x(t+1) = Ax(t) + \omega(t), \qquad x(0) \sim \mathcal{N}(0, \Sigma_0), \quad \forall t \in \mathbb{N},$$

where $\omega(t)$ are i.i.d. according to $\mathcal{N}(0, \Sigma_\omega)$ and are independent of $x(0)$. The $\omega(t)$'s have zero means for all $t \in \mathbb{T}$, and thus the mean of the system is 0 for all $t \in \mathbb{N}$. The covariances $\Sigma(t)$'s satisfy the recursion,

$$\Sigma(t+1) = A\Sigma(t)A^\top + \Sigma_\omega, \quad \forall t \in \mathbb{N}, \tag{31}$$

which converges to the steady-state covariance if and only if A is stable. In this experiment, we consider a 50-dimensional system, i.e., $x(t) \in \mathbb{R}^{50}$ for all $t \in \mathbb{N}$, where the matrix $A \in \mathbb{R}^{50\times 50}$ is generated via random.rand() in Python. The maximum eigenvalue of A is 1.25, and thus the iterates in (31), diverge. We take $\Sigma_0 = 0.2$ Id and $\Sigma_\omega = 0.01$ Id. In order to learn the CME operator, we collected $|\mathcal{D}| = 2,500$ samples uniformly distributed on $[0,1]^{50}$ and propagated them one step forward with a sampling interval of 0.01 s. The kernel function is chosen as a combination of three Gaussian kernels $\kappa(x,y) = \sum_{i=1}^{3} \eta_i \exp\left(\frac{-\|x-y\|^2}{2\times\sigma_i^2}\right)$ where $(\eta_1, \eta_2, \eta_3) = (0.1, 0.8, 0.1)$ and $(\sigma_1, \sigma_2, \sigma_3) = (4, 4.8, 11)$. We then set $\gamma = 0.52$ to get the sparse dictionary $\widehat{\mathcal{D}}$ with $|\widehat{\mathcal{D}}| = 2278$. We applied Algorithm 1 by setting the regularization parameter as $\varepsilon = 1e^{-13} \times |\widehat{\mathcal{D}}|^{-0.2}$. We compared estimates of moments with the analytic solutions. Figure 2 indicates that our algorithm is able to form accurate estimates of the true moments that can be calculated analytically.

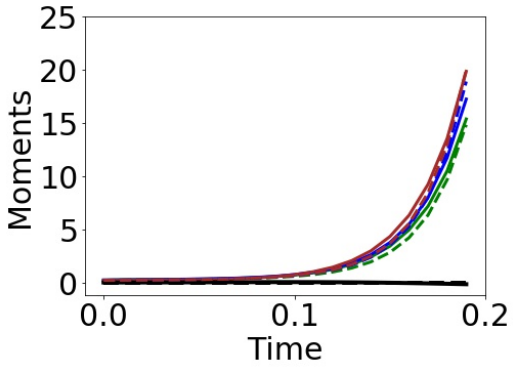


Figure 2: Approximations of the mean (—) and second-order moments (—) (—) (—) along three exemplary dimensions of the linear system with Gaussian noise. Dashed lines are those obtained from analytic solutions.

4.2. A Simple 2-D System

Consider a 2-dimensional nonlinear dynamical system

$$\dot{x}_1 = x_2, \quad \dot{x}_2 = -\frac{3}{2}x_1 - x_2 + \frac{x_2^3}{9}, \tag{32}$$

which admits a stable equilibrium at $(0, 0)$. As shown in Figure 3a, trajectories starting from the initial uncertainty set $(-1.5, -1.1) \times (0.4, 0.8)$ converge to the stable equilibrium point $(0, 0)$. In practice, observations are often corrupted by noise. To this end, we add Gaussian noise with zero mean and standard deviation of $1e^{-3}$ to sampled data points.

In order to construct $\widehat{\mathcal{U}}_\varepsilon$ defined in (23) [4], we randomly selected $1,000$ initial points in a circle around the origin with a radius of 3 and numerically integrated 9 steps forward in Python with odeint-solver using $\Delta t = 0.1$ s. The kernel function is a combination of two Gaussian kernels $\kappa(x, y) = \sum_{i=1}^{2} \eta_i \exp\left(\frac{-\|x-y\|^2}{2\times\sigma_i^2}\right)$ where $(\eta_1, \eta_2) = (0.5, 0.5)$ and $(\sigma_1, \sigma_2) = (0.7, 0.18)$. We then pruned the dataset comprising $9,000$ sample pairs to obtain $|\widehat{\mathcal{D}}| = 2186$ using $\gamma = 0.99$ and use $\varepsilon = 1e^{-13} \times |\widehat{\mathcal{D}}|^{-0.2}$. Figure 3b shows the comparison between the estimated moments obtained from Algorithm 1 and the Monte Carlo method. The convergence of all moments to zero indicates that all trajectories converge to the stable equilibrium point eventually.

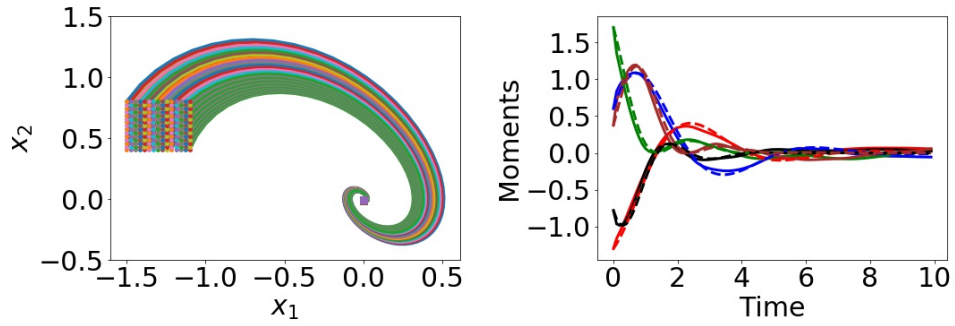
4.3. Genetic Bi-stable Toggle

We next consider the kinetics of the concentration of two proteins that inhibit each other, also known as the genetic toggle switch per Gardner et al. [56], described by

$$\dot{x}_1 = \frac{1}{1 + x_2^3.55} - 0.5x_1, \quad \dot{x}_2 = \frac{1}{1 + x_1^3.53} - 0.5x_2. \tag{33}$$

As shown in Figure 4a, the system admits two equilibrium points–$(0.16, 2)$ and $(0.161, 0.2)$–with complementary regions of attraction. We consider two initial uncertainty regions with different shapes: (a) $\mathcal{S}_c$–a circle centered at $(0.4, 0.8)$ with radius 0.2 and (b) $\mathcal{S}_s$–a square $(1.2, 1.4) \times (0.5, 0.7)$. Figure

[4]For the case of a deterministic dynamical system, we consider the embedding of the Dirac delta function. Under the assumption that the kernel function is bounded, such embedding is well-defined and we use $\widehat{\mathcal{U}}_\varepsilon$ as its approximation.



(a) (b)

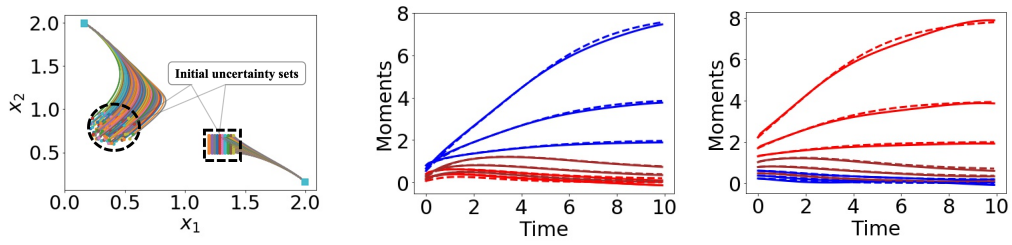
Figure 3: (a) Trajectories starting from uncertainty set $(-1.5, -1.1) \times (0.4, 0.8)$ that converge to the stable equilibrium point $(0, 0)$. (b) Estimated moments up to order 2 using the proposed algorithm (solid lines) versus the Monte Carlo method (dashed lines). $\mathbb{E}[x_1]$(—),$\mathbb{E}[x_2]$(—),$\mathbb{E}[x_1^2]$(—),$\mathbb{E}[x_1 x_2]$(—),$\mathbb{E}[x_2^2]$(—) .

4a demonstrates that samples starting from those two initial uncertainty sets converge to $(0.16, 2)$ and $(0.161, 0.2)$. We also consider the case where observations are corrupted by additive Gaussian noise with zero mean and standard deviation of $1e^{-4}$.

To compute $\widehat{\mathcal{U}}_\varepsilon$, $1,600$ initial points were selected over $[0, 0.25] \times [0, 0.25]$. We then numerically integrated 9 steps forward with a time interval of $\Delta t = 0.1$s. The kernel function is a combination of two Gaussian kernels $\kappa(x, y) = \sum_{i=1}^{2} \eta_i \exp\left(\frac{-\|x-y\|^2}{2\times\sigma_i^2}\right)$ where $(\eta_1, \eta_2) = (0.55, 0.45)$ and $(\sigma_1, \sigma_2) = (0.475, 1)$. Next, we pruned the original dataset consisting of $14,400$ samples with $\gamma = 0.99$ to obtain a sparse dictionary with $|\widehat{\mathcal{D}}| = 1463$ and use $\varepsilon = 1e^{-13} \times |\widehat{\mathcal{D}}|^{-0.2}$. Following Algorithm 1, the true and estimated moments are plotted in Figure 4. Comparing Figure 4b with 4c suggests that different moments dominate the plots as time advances. In the case of Figure 4b, the moments that are expressions of only x_2 rise with time while the moments concerning x_1 decay with time. Those trends imply that all trajectories starting from $\mathcal{S}_c$ converge to $(0.16, 2)$. The opposite is true in Figure 4c.

5. Applications to Power Systems

In this section, we apply our framework to propagating uncertainty in example power systems. We illustrate that our proposed data-driven method



(a) (b) (c)

Figure 4: Starting from two initial uncertainty sets, the estimated moments up to order 3 are plotted using the proposed algorithm (solid lines) and the Monte Carlo method (dashed lines). In (a), moments that are dominated by x_2, i.e., $\mathbb{E}[x_2], \mathbb{E}[x_2^2], \mathbb{E}[x_2^3], \mathbb{E}[x_1 x_2^2]$ (—) rise with time while moments dominated by x_1, i.e., $\mathbb{E}[x_1], \mathbb{E}[x_1^2], \mathbb{E}[x_1^3], \mathbb{E}[x_1^2 x_2]$(—) as well as $\mathbb{E}[x_1 x_2]$(—), decay with time. The reverse is true in (b).

is faster than the Monte Carlo method. All experiments were run on a MacBook Pro with Apple M1 pro chip.

5.1. Single Machine Infinite Bus System (SMIB)

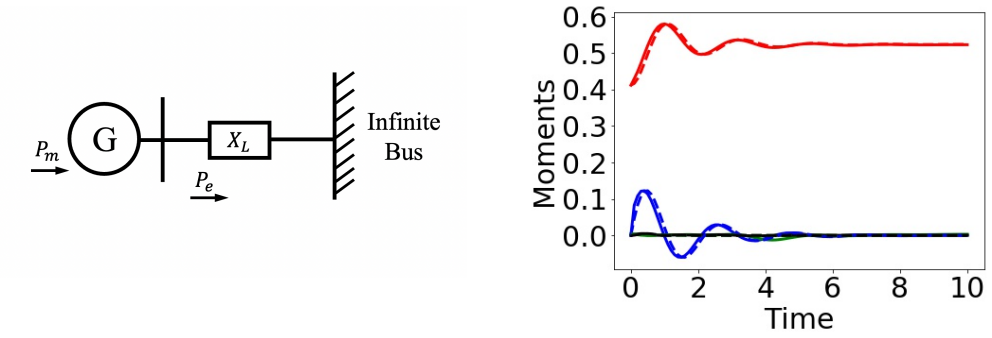
Consider the dynamical system of SMIB, described by

$$\dot{\delta} = \omega, \quad \dot{\omega} = -D\omega + P_m - P_e \sin(\delta), \tag{34}$$

with $D = 1.3$, $P_m = 5$ and $P_e = 10$. The stable equilibrium of this system is at $[0.53\ rad, 0]$. We are interested in uncertainties in the rotor angles of the generator with 0.3 rad $\leq \delta \leq \pi/6$ rad.

To build a sparse representation of the underlying dynamics, we first sampled 900 initial points $x = [\delta, \omega]$ that are uniformly distributed over $[\delta, \omega] \in [-4, 4] \times [-8, 8]$. We then collected 10 points along each trajectory with sampling interval $\Delta t = 0.1$. The kernel function is a combination of three Gaussian kernels $\kappa(x, y) = \sum_{i=1}^{3} \eta_i \exp\left(\frac{-\|x-y\|^2}{2\times\sigma_i^2}\right)$ where $(\eta_1, \eta_2, \eta_3) = (0.1, 0.8, 0.1)$ and $(\sigma_1, \sigma_2, \sigma_3) = (0.12, 0.56, 1)$. We pruned the dataset utilize $\gamma = 0.9$ to get $\left|\widehat{\mathcal{D}}\right| = 2081$ and use $\varepsilon = 1e^{-13} \times |\widehat{\mathcal{D}}|^{-0.2}$. $\widehat{\mu}_t$ and moments are then computed based on $\widehat{\mathcal{U}}_\varepsilon$. The comparison with results from the Monte Carlo-based moment propagation in Figure 5b reveals that Algorithm 1 achieves very similar accuracy as the Monte-Carlo method, and that the system returns to the

nominal operating condition with time as the variance is damped down to zero. In particular, with learned $\widehat{\mathcal{U}}_\varepsilon$, the computation time for a time horizon of 10s takes only 2.2s. By contrast, the Monte Carlo simulation takes 20.4s as it is typically more data intensive.



(a) (b)

Figure 5: (a) One-line diagram of SMIB. (b) Propagation of mean $\mathbb{E}[\delta]$(—), $\mathbb{E}[\omega]$(—) and variance $\mathrm{Var}[\delta]$(—), $\mathrm{Var}[\omega]$(—) using the Algorithm 1 (solid lines) and Monte Carlo method (dashed lines).

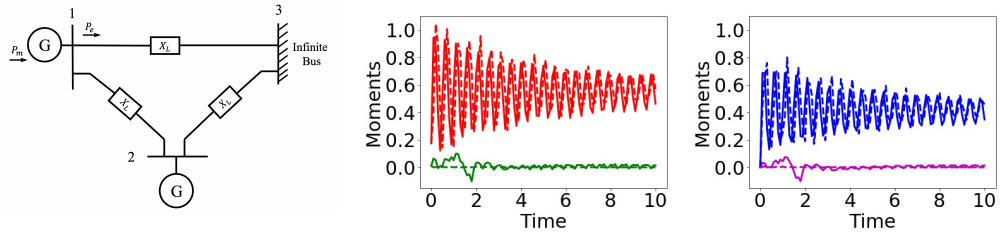
5.2. Two Machine Infinite Bus System (TMIB)

Next, consider the TMIB described by

$$\begin{aligned}\dot{\delta}_i =&\omega_i,\\ M_i\dot{\omega}_1 =&P_{m,i} - D_i\omega_i - \frac{E_i}{x_i}\sin(\delta_i) - \frac{E_1E_2}{x_{12}}\sin(\delta_i-\delta_j),\end{aligned}$$

for $i = 1, 2$. The uncertainties in the initial conditions are in rotor angles of generator one with 0 rad $\leq \delta_1 \leq$ 0.35 rad. We scale ω_1, ω_2 by a factor of 1/8 and utilize a combination of three Gaussian kernels $\kappa(x, y) = \sum_{i=1}^{3} \eta_i \exp\left(\frac{-\|x-y\|^2}{2\times\sigma_i^2}\right)$ where $(\eta_1, \eta_2, \eta_3) = (0.45, 0.35, 0.2)$ and $(\sigma_1, \sigma_2, \sigma_3) = (0.15, 0.38, 0.75)$. To compute $\widehat{\mathcal{U}}_\varepsilon$, we collected 5,000 sample pairs from 500 trajectories with 10 evaluations along each and constructed a sparse dictionary with $|\widehat{\mathcal{D}}| = 2653$ and use $\varepsilon = 1e^{-13} \times |\widehat{\mathcal{D}}|^{-0.2}$. $\widehat{\mathcal{U}}_\varepsilon$ is then used to estimates $\widehat{\mu}_t$ for $t \in \mathbb{T}$ following line 4-6 in Algorithm 1. Figure 6 plots the mean and variance of rotor angles obtained from Algorithm 1 and the Monte Carlo method. It can be seen that the uncertainty of δ_1 is propagated to δ_2

as its mean and variance increase from 0 before decaying. We also observe that eventually, $\mathbb{V}[\delta_1], \mathbb{V}[\delta_2]$ stay close to 0 eventually, indicating that the system converges to the nominal operating point.



(a) (b) Generator 1 (c) Generator 2

Figure 6: (a) One-line diagram of TMIB. (b)(c) Plots of mean $\mathbb{E}[\delta_1]$(—),$\mathbb{E}[\delta_2]$(—) and variance $\mathrm{Var}[\delta_1]$(—),$\mathrm{Var}[\delta_2]$(—) of two generators using Algorithm 1 (solid lines) and Monte Carlo method (dashed lines).

For this particular example, constructing $\widehat{\mathcal{U}}_\varepsilon$ took around 182.07 seconds. Computing propagated moments for a time horizon of 10s with a learned sparse model took merely 3.65 s since only algebraic operations are involved. By contrast, the Monte Carlo simulation took around 1020.33 second, which is even beyond the time horizon of interest. In conclusion, the proposed method is more computationally efficient for the propagation of initial uncertainty.

Our experiments suggest $\varepsilon = 1e^{-13} \times |\widehat{\mathcal{D}}|^{-0.2}$ as a good thumb rule for the choice of the regularization parameter. A more comprehensive empirical work is needed to test the efficacy of such a choice. In our experiments, we utilized Gaussian kernels with different widths. According to Sriperumbudur et al. [57], the Gaussian kernel is characteristic, and thus, the embedding can preserve all information about the distribution. Automating the process of choosing an appropriate kernel is an interesting direction for future work. As for the coherence parameter, upon decreasing the value of γ, we obtain a less coherent dictionary $\widehat{\mathcal{D}}$ with fewer elements. For a kernel function with $B_\kappa = 1$, one typically chooses γ between 0.5 and 1. We refer interested readers to Hou et al. [38] for a detailed discussion of the practical benefits of sparsification and the role of γ.

6. Conclusions

In this paper, we provided an algorithm to propagate uncertainty in initial conditions through unknown system dynamics in RKHS. A sparse representation of the dynamical system is learned through the CME operator. We have provided sample complexity bounds for approximations of embedded uncertainty. Five exemplary numerical experiments confirmed the effectiveness of our approach. Scaling the proposed framework to larger power system examples is an interesting direction for future work.

Acknowledgements

This work was supported by the NSF-EPCN-2031570 grant and NSF-CPS-2038775 grant.

References

[1] H. Choi, P. J. Seiler, S. V. Dhople, Propagating uncertainty in power-system dae models with semidefinite programming, IEEE Transactions on Power Systems 32 (2016) 3146–3156.

[2] Y. C. Chen, A. D. Domínguez-García, Assessing the impact of wind variability on power system small-signal reachability, in: 2011 44th Hawaii International Conference on System Sciences, IEEE, 2011, pp. 1–8.

[3] H. N. V. Pico, D. C. Aliprantis, E. C. Hoff, Reachability analysis of power system frequency dynamics with new high-capacity hvac and hvdc transmission lines, in: 2013 IREP Symposium Bulk Power System Dynamics and Control-IX Optimization, Security and Control of the Emerging Power Grid, IEEE, 2013, pp. 1–9.

[4] X. Jiang, A. D. Domínguez-García, A zonotope-based method for capturing the effect of variable generation on the power flow, in: 2014 North American Power Symposium (NAPS), IEEE, 2014, pp. 1–6.

[5] J. H. Halton, A retrospective and prospective survey of the monte carlo method, Siam review 12 (1970) 1–63.

[6] K. M. Hanson, A framework for assessing uncertainties in simulation predictions, Physica D: Nonlinear Phenomena 133 (1999) 179–188.

[7] J. C. Helton, Treatment of uncertainty in performance assessments for complex systems, Risk analysis 14 (1994) 483–511.

[8] I. Mezic, T. Runolfsson, Uncertainty analysis of complex dynamical systems, in: Proceedings of the 2004 American Control Conference, volume 3, IEEE, 2004, pp. 2659–2664.

[9] Y. Xu, L. Mili, A. Sandu, M. R. von Spakovsky, J. Zhao, Propagating uncertainty in power system dynamic simulations using polynomial chaos, IEEE Transactions on Power Systems 34 (2018) 338–348.

[10] A. R. R. Matavalam, U. Vaidya, V. Ajjarapu, Data-driven approach for uncertainty propagation and reachability analysis in dynamical systems, in: 2020 American Control Conference (ACC), IEEE, 2020, pp. 3393–3398.

[11] R. Ghanem, J. Red-Horse, Propagation of probabilistic uncertainty in complex physical systems using a stochastic finite element approach, Physica D: Nonlinear Phenomena 133 (1999) 137–144.

[12] R. G. Ghanem, P. D. Spanos, Stochastic finite elements: a spectral approach, Courier Corporation, 2003.

[13] A. R. R. Matavalam, U. Vaidya, V. Ajjarapu, Propagating uncertainty in power system initial conditions using data-driven linear operators, IEEE Transactions on Power Systems 37 (2022) 4125–4128.

[14] A. Lasota, M. C. Mackey, Chaos, fractals, and noise: stochastic aspects of dynamics, volume 97, Springer Science & Business Media, 1998.

[15] I. Mezić, A. Banaszuk, Comparison of systems with complex behavior, Physica D: Nonlinear Phenomena 197 (2004) 101–133.

[16] I. Mezić, Spectral properties of dynamical systems, model reduction and decompositions, Nonlinear Dynamics 41 (2005) 309–325.

[17] A. Mauroy, I. Mezić, Global stability analysis using the eigenfunctions of the koopman operator, IEEE Transactions on Automatic Control 61 (2016) 3356–3369.

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

[19] S. Klus, P. Koltai, C. Schütte, On the numerical approximation of the perron-frobenius and koopman operator, arXiv preprint arXiv:1512.05997 (2015).

[20] M. Korda, I. Mezić, On convergence of extended dynamic mode decomposition to the koopman operator, Journal of Nonlinear Science 28 (2018) 687–710.

[21] Q. Li, F. Dietrich, E. M. Bollt, I. G. Kevrekidis, Extended dynamic mode decomposition with dictionary learning: A data-driven adaptive spectral decomposition of the koopman operator, Chaos: An Interdisciplinary Journal of Nonlinear Science 27 (2017) 103111.

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

[23] N. Takeishi, Y. Kawahara, T. Yairi, Learning koopman invariant subspaces for dynamic mode decomposition, Advances in neural information processing systems 30 (2017).

[24] B. Lusch, J. N. Kutz, S. L. Brunton, Deep learning for universal linear embeddings of nonlinear dynamics, Nature communications 9 (2018) 4950.

[25] C. Wehmeyer, F. Noé, Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics, The Journal of chemical physics 148 (2018).

[26] S. E. Otto, C. W. Rowley, Linearly recurrent autoencoder networks for learning dynamics, SIAM Journal on Applied Dynamical Systems 18 (2019) 558–593.

[27] M. O. Williams, C. W. Rowley, I. G. Kevrekidis, A kernel-based approach to data-driven koopman spectral analysis, Journal of Computational Dynamics (2014).

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

[29] S. Klus, I. Schuster, K. Muandet, Eigendecompositions of transfer operators in reproducing kernel hilbert spaces, Journal of Nonlinear Science 30 (2020) 283–315.

[30] B. Hou, S. Sanjari, N. Dahlin, S. Bose, U. Vaidya, Sparse learning of dynamical systems in rkhs: An operator-theoretic approach, in: International Conference on Machine Learning, PMLR, 2023, pp. 13325–13352.

[31] V. Kostic, P. Novelli, A. Maurer, C. Ciliberto, L. Rosasco, M. Pontil, Learning dynamical systems via koopman operator regression in reproducing kernel hilbert spaces, Advances in Neural Information Processing Systems 35 (2022) 4017–4031.

[32] G. Meanti, A. Chatalic, V. R. Kostic, P. Novelli, M. Pontil, L. Rosasco, Estimating koopman operators with sketching to provably learn large scale dynamical systems, arXiv preprint arXiv:2306.04520 (2023).

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

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

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

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

[37] B. Hou, S. Sanjari, N. Dahlin, S. Bose, Compressed decentralized learning of conditional mean embedding operators in reproducing kernel hilbert spaces, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, 2023, pp. 7902–7909.

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

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

[40] J. Kivinen, A. J. Smola, R. C. Williamson, Online learning with kernels, IEEE transactions on signal processing 52 (2004) 2165–2176.

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

[42] A. Rahimi, B. Recht, Random features for large-scale kernel machines, Advances in neural information processing systems 20 (2007).

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

[44] A. Chatalic, N. Schreuder, L. Rosasco, A. Rudi, Nyström kernel mean embeddings, in: International Conference on Machine Learning, PMLR, 2022, pp. 3006–3024.

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

[46] J.-J. Zhu, K. Muandet, M. Diehl, B. Schölkopf, A new distribution-free concept for representing, comparing, and propagating uncertainty in dynamical systems with kernel probabilistic programming, IFAC-PapersOnLine 53 (2020) 7240–7247.

[47] L. Song, A. Gretton, C. Guestrin, Nonparametric tree graphical models, in: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings, 2010, pp. 765–772.

[48] S. P. Nandanoori, S. Sinha, E. Yeung, Data-driven operator theoretic methods for global phase space learning, in: 2020 American Control Conference (ACC), IEEE, 2020, pp. 4551–4557.

[49] J. Park, K. Muandet, A measure-theoretic approach to kernel conditional mean embeddings, Advances in Neural InformationProcessing Systems 33 (2020) 21247–21259.

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

[51] Z. Li, D. Meunier, M. Mollenhauer, A. Gretton, Optimal rates for regularized conditional mean embedding learning, Advances in Neural Information Processing Systems 35 (2022) 4433–4445.

[52] M. Dellnitz, O. Junge, On the approximation of complicated dynamical behavior, SIAM Journal on Numerical Analysis 36 (1999) 491–515.

[53] Y. Engel, S. Mannor, R. Meir, The kernel recursive least-squares algorithm, IEEE Transactions on signal processing 52 (2004) 2275–2285.

[54] A. Gretton, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, K. Fukumizu, B. K. Sriperumbudur, Optimal kernel choice for large-scale two-sample tests, Advances in neural information processing systems 25 (2012).

[55] K. Fukumizu, A. Gretton, G. Lanckriet, B. Schölkopf, B. K. Sriperumbudur, Kernel choice and classifiability for rkhs embeddings of probability distributions, Advances in neural information processing systems 22 (2009).

[56] T. S. Gardner, C. R. Cantor, J. J. Collins, Construction of a genetic toggle switch in escherichia coli, Nature 403 (2000) 339–342.

[57] B. K. Sriperumbudur, K. Fukumizu, G. R. Lanckriet, Universality, characteristic kernels and rkhs embedding of measures., Journal of Machine Learning Research 12 (2011).

[58] J.-P. Aubin, Applied functional analysis, John Wiley & Sons, 2011.

Appendix A. Proof of Theorem 1

We start with the following lemma that provides a uniform bound on the operator norm of $\mathcal{U}_\varepsilon$ and its sparse estimate $\widehat{\mathcal{U}}_\varepsilon$.

Lemma 1. *Under assumptions of Theorem 1, $\mathcal{U}_\varepsilon$ defined in (11) and $\widehat{\mathcal{U}}_\varepsilon$ defined in (23) satisfies $\|\mathcal{U}_\varepsilon\| \leq B_\kappa/\varepsilon$ and $\left\|\widehat{\mathcal{U}}_\varepsilon\right\| \leq B_\kappa/\varepsilon$.*

Proof. To prove the first claim, notice that by definition,

$$\begin{aligned} \|\mathcal{U}_\varepsilon\| &= \left\|C_{X^+X}\left(C_{XX}+\varepsilon\ \mathrm{Id}\right)^{-1}\right\| \\ &\leq \|C_{X^+X}\| \left\|\left(C_{XX}+\varepsilon\ \mathrm{Id}\right)^{-1}\right\| \\ &\leq \frac{1}{\varepsilon}\|C_{X^+X}\|, \end{aligned} \tag{A.1}$$

where the last line follows from the fact that C_{XX} is positive semi-definite and self-adjoint, which implies $\left\|\left(C_{XX}+\varepsilon\ \mathrm{Id}\right)^{-1}\right\| \leq 1/\varepsilon$. In order to bound $\|C_{X^+X}\|$, note that C_{X^+X} is a Hilbert-Schmidt operator and that the space of Hilbert-Schmidt operator from $\mathcal{H}$ to $\mathcal{H}$ is isometric isomorphism to the tensor product Hilbert space $\mathcal{H}\otimes\mathcal{H}$ Aubin [58]. Together with the fact that the operator norm is dominated by the Hilbert-Schmidt norm, we have

$$\begin{aligned} \|C_{X^+X}\|^2 \leq& \|C_{X^+X}\|_{\mathrm{HS}}^2 \\ =& \left\|\mathbb{E}_{X^+X}\left[\phi(X^+)\otimes\phi(X)\right]\right\|_{\mathcal{H}\otimes\mathcal{H}}^2 \\ =& \left\langle \mathbb{E}_{X^+X}\left[\phi(X^+)\otimes\phi(X)\right], \mathbb{E}_{X'^+X'}\left[\phi(X'^+)\otimes\phi(X')\right]\right\rangle_{\mathcal{H}\otimes\mathcal{H}} \\ =& \mathbb{E}_{X^+X,X'^+X'}\left[\left\langle \phi(X^+)\otimes\phi(X), \phi(X'^+)\otimes\phi(X')\right\rangle_{\mathcal{H}\otimes\mathcal{H}}\right] \\ =& \mathbb{E}_{X^+X,X'^+X'}\left[\left\langle \phi(X^+), \phi(X'^+)\right\rangle_{\mathcal{H}} \left\langle\phi(X),\phi(X')\right\rangle_{\mathcal{H}}\right] \\ =& \mathbb{E}_{X^+X,X'^+X'}\left[\kappa(X^+,X'^+)\kappa(X,X')\right] \\ \leq& B_\kappa^2, \end{aligned} \tag{A.2}$$

where (X', X'^+) is an independent copy of (X, X^+) and the last line follows from boundedness of κ and the fact that $\sup_{x,x'\in\mathbb{X}}|\kappa(x,x')| = \sup_{x\in\mathbb{X}}\kappa(x,x)$ per Steinwart and Christmann [33, Section 4.3].

Likewise, we have

$$\begin{aligned}\left\|\widehat{\mathcal{U}}_\varepsilon\right\| &= \left\|\widehat{C}_{X^+X}\left(\widehat{C}_{XX}+\varepsilon\ \mathrm{Id}\right)^{-1}\right\| \\ &\le \left\|\widehat{C}_{X^+X}\right\|\left\|\left(\widehat{C}_{XX}+\varepsilon\ \mathrm{Id}\right)^{-1}\right\| \\ &\le \frac{1}{\varepsilon}\left\|\widehat{C}_{X^+X}\right\|.\end{aligned} \tag{A.3}$$

By definition, $\widehat{C}_{X^+X}$ defined in (19),(20) is the projection of $\widetilde{C}_{X^+X}$ defined in (16) onto the closed subspace $\{\varphi(x_i^+, x_i) : i \in \mathcal{I}\}$ of $\mathcal{H}\otimes\mathcal{H}$. Let $\Pi_{\widehat{\mathcal{D}}}$ be the (linear) projection operator and we have

$$\left\|\widehat{C}_{X^+X}\right\| \le \left\|\widehat{C}_{X^+X}\right\|_{\mathrm{HS}} = \left\|\Pi_{\widehat{\mathcal{D}}}\left(\widetilde{C}_{X^+X}\right)\right\|_{\mathrm{HS}} \le \left\|\widetilde{C}_{X^+X}\right\|_{\mathrm{HS}} \tag{A.4}$$

On the other hand,

$$\begin{aligned}\left\|\widetilde{C}_{X^+X}\right\|_{\mathrm{HS}}^2 =& \left\|\frac{1}{M}\sum_{i=1}^M \phi(x_i^+)\otimes\phi(x_i)\right\|_{\mathcal{H}\otimes\mathcal{H}}^2 \\ =& \left\langle \frac{1}{M}\sum_{i=1}^M \phi(x_i^+)\otimes\phi(x_i), \frac{1}{M}\sum_{j=1}^M \phi(x_j^+)\otimes\phi(x_j)\right\rangle_{\mathcal{H}\otimes\mathcal{H}} \\ =& \frac{1}{M^2}\sum_{i,j=1}^M \left\langle \phi(x_i^+)\otimes\phi(x_i), \phi(x_j^+)\otimes\phi(x_j)\right\rangle_{\mathcal{H}\otimes\mathcal{H}} \\ =& \frac{1}{M^2}\sum_{i,j=1}^M \left[\left\langle \phi(x_i^+), \phi(x_j{}^+)\right\rangle_{\mathcal{H}} \left\langle \phi(x_i), \phi(x_j)\right\rangle_{\mathcal{H}}\right] \\ =& \frac{1}{M^2}\sum_{i,j=1}^M \kappa(x_i^+, x_j{}^+)\kappa(x_i, x_j) \\ \le& B_\kappa^2,\end{aligned} \tag{A.5}$$

Combining eqs. (A.1) to (A.5) gives

$$\|\mathcal{U}_\varepsilon\| \le B_u, \quad \left\|\widehat{\mathcal{U}}_\varepsilon\right\| \le B_u, \quad B_u := \frac{B_\kappa}{\varepsilon}. \tag{A.6}$$

□

We now return to the proof of Theorem 1. At time $t \in \mathbb{T}$, we have

$$\begin{aligned}
\|\mu_t - \widehat{\mu}_t\|_{\mathcal{H}} &= \left\|\mathcal{U}_\varepsilon^t \mu_0 - \widehat{\mathcal{U}}_\varepsilon^t \widehat{\mu}_0\right\|_{\mathcal{H}} \\
&= \left\|\mathcal{U}_\varepsilon^t \mu_0 - \mathcal{U}_\varepsilon^t \widehat{\mu}_0 + \mathcal{U}_\varepsilon^t \widehat{\mu}_0 - \widehat{\mathcal{U}}_\varepsilon^t \widehat{\mu}_0\right\|_{\mathcal{H}} \\
&\overset{(a)}{\leq} \left\|\mathcal{U}_\varepsilon^t \left(\mu_0 - \widehat{\mu}_0\right)\right\|_{\mathcal{H}} + \left\|\left(\mathcal{U}_\varepsilon^t - \widehat{\mathcal{U}}_\varepsilon^t\right) \widehat{\mu}_0\right\|_{\mathcal{H}} \\
&\leq \left\|\mathcal{U}_\varepsilon^t\right\| \|\mu_0 - \widehat{\mu}_0\|_{\mathcal{H}} + \left\|\mathcal{U}_\varepsilon^t - \widehat{\mathcal{U}}_\varepsilon^t\right\| \|\widehat{\mu}_0\|_{\mathcal{H}} \\
&\overset{(b)}{\leq} \|\mathcal{U}_\varepsilon\|^t \|\mu_0 - \widehat{\mu}_0\|_{\mathcal{H}} + \left\|\mathcal{U}_\varepsilon^t - \widehat{\mathcal{U}}_\varepsilon^t\right\| \sqrt{B_\kappa} \\
&\overset{(c)}{\leq} B_u^t \|\mu_0 - \widehat{\mu}_0\|_{\mathcal{H}} + \left\|\mathcal{U}_\varepsilon^t - \widehat{\mathcal{U}}_\varepsilon^t\right\| \sqrt{B_\kappa},
\end{aligned} \tag{A.7}$$

where in (a), we break down the error into two terms: the first one encodes propagation of approximation error in the initial estimate through iterations and the second one captures the error at time t when starting from the same initial empirical estimator. Line (b) follows from submultiplicative of the operator norm and (c) follows from Lemma 1

In order to bound $\left\|\mathcal{U}_\varepsilon^t - \widehat{\mathcal{U}}_\varepsilon^t\right\|$, we have

$$\begin{aligned}
\left\|\mathcal{U}_\varepsilon^t - \widehat{\mathcal{U}}_\varepsilon^t\right\| &= \left\|\left(\mathcal{U}_\varepsilon - \widehat{\mathcal{U}}_\varepsilon\right)\left(\sum_{k=1}^{t} \mathcal{U}_\varepsilon^{t-k} \widehat{\mathcal{U}}_\varepsilon^{k-1}\right)\right\| \\
&\leq \left\|\mathcal{U}_\varepsilon - \widehat{\mathcal{U}}_\varepsilon\right\| \left(\sum_{k=1}^{t} \left\|\mathcal{U}_\varepsilon^{t-k} \widehat{\mathcal{U}}_\varepsilon^{k-1}\right\|\right) \\
&\leq \left\|\mathcal{U}_\varepsilon - \widehat{\mathcal{U}}_\varepsilon\right\| \left(\sum_{k=1}^{t} \|\mathcal{U}_\varepsilon\|^{t-k} \left\|\widehat{\mathcal{U}}_\varepsilon\right\|^{k-1}\right) \\
&\leq \left\|\mathcal{U}_\varepsilon - \widehat{\mathcal{U}}_\varepsilon\right\| \left(t B_u^{t-1}\right),
\end{aligned} \tag{A.8}$$

where we apply Lemma 1 to get the last line. Plugging into (A.7) gives

$$\|\mu_t - \widehat{\mu}_t\|_{\mathcal{H}} \leq B_u^t \|\mu_0 - \widehat{\mu}_0\|_{\mathcal{H}} + t B_u^{t-1} \sqrt{B_\kappa} \left\|\mathcal{U}_\varepsilon - \widehat{\mathcal{U}}_\varepsilon\right\|. \tag{A.9}$$

Next, we apply Muandet et al. [36, Theorem 3.4] and Hou et al. [30, Theorem 1] to bound $\|\mu_0 - \widehat{\mu}_0\|_{\mathcal{H}}$ and $\left\|\mathcal{U}_\varepsilon - \widehat{\mathcal{U}}_\varepsilon\right\|$, respectively. We then conclude

that with probability at least $1-\delta$,

$$\begin{aligned}
\|\mu_t - \widehat{\mu}_t\|_{\mathcal{H}} \leq & B_u^t \frac{\sqrt{B_\kappa}}{\sqrt{N}} \Xi\left(\delta/2\right) + t B_u^{t-1} B_\kappa^{3/2} \psi\left(M, \gamma; \delta/4\right) \left(\frac{1}{\varepsilon} + \frac{\|C_{X^+X}\|}{\varepsilon^2}\right) \\
= & \left(\frac{B_\kappa}{\varepsilon}\right)^t \frac{\sqrt{B_\kappa}}{\sqrt{N}} \Xi\left(\delta/2\right) + t \left(\frac{B_\kappa}{\varepsilon}\right)^{t-1} B_\kappa^{3/2} \psi\left(M, \gamma; \delta/4\right) \left(\frac{1}{\varepsilon} + \frac{\|C_{X^+X}\|}{\varepsilon^2}\right) \\
= & {B_\kappa}^t \left(\frac{1}{\varepsilon^t} \frac{\sqrt{B_\kappa}}{\sqrt{N}} \Xi\left(\delta/2\right) + t \frac{1}{\varepsilon^{t-1}} B_\kappa^{1/2} \psi\left(M, \gamma; \delta/4\right) \left(\frac{1}{\varepsilon} + \frac{\|C_{X^+X}\|}{\varepsilon^2}\right)\right) \\
= & {B_\kappa}^t \left(\frac{1}{\sqrt{N}} \Xi\left(\delta/2\right) \mathcal{O}\left(\varepsilon^{-t}\right) + t \psi\left(M, \gamma; \delta/4\right) \mathcal{O}\left(\varepsilon^{-(t+1)}\right)\right),
\end{aligned} \tag{A.10}$$

where ψ is defined in (30).