Consistent Long-Term Forecasting of Ergodic Dynamical Systems

Vladimir Kostic, Prune Inzerili, Karim Lounici, Pietro Novelli, Massimiliano Pontil

Consistent Long-Term Forecasting of Ergodic Dynamical Systems

Prune Inzerili [1] **Vladimir Kostic** [2 3] **Karim Lounici** [1] **Pietro Novelli** [2] **Massimiliano Pontil** [2 4]

Abstract

We study the problem of forecasting the evolution of a function of the state (observable) of a discrete ergodic dynamical system over multiple time steps. The elegant theory of Koopman and transfer operators can be used to evolve any such function forward in time. However, their estimators are usually unreliable in long-term forecasting. We show how classical techniques of eigenvalue deflation from operator theory and feature centering from statistics can be exploited to enhance standard estimators. We develop a novel technique to derive high probability bounds on powers of empirical estimators. Our approach, rooted in the stability *theory of non-normal operators*, allows us to establish uniform in time bounds for the forecasting error, which hold even on *infinite time horizons*. We further show that our approach can be seamlessly employed to forecast future state distributions from an initial one, with provably uniform error bounds. Numerical experiments illustrate the advantages of our approach in practice.

1. Introduction

Dynamical systems offer a mathematical framework to describe the evolution of state variables over time. In many applications, these models, often represented by unknown nonlinear differential equations (ordinary or partial, and possibly stochastic), necessitate the use of data-driven techniques for characterizing the dynamical system and forecasting future states. This task has garnered substantial interest in recent decades due its application in many fields, including energy forecasting (Mohan et al., 2018), epidemiology (Proctor & Eckhoff, 2015), finance (Pascucci, 2011), atomistic simulations (Schütte et al., 2001), fluid dynamics (Mezić, 2013), weather and climate forecasting (Scher, 2018), and many more.

Particular emphasis is placed on long-term forecasting, which, given any initial state of the system, aims to predict how it (or a given statistics thereof) will evolve over time, until a long-term horizon. The accuracy of long-term forecasting is of utmost importance for effective strategic planning and early warning systems. However structured data modalities, an increasing volume of observations, and highly non-linear relationships among covariates pose significant challenges to current approaches. In this work, we specifically address the problem of long-term forecasting of ergodic dynamical systems, whose states converge to an unknown but invariant distribution over time.

The Koopman[1] operator regression (KOR) framework to learn dynamical systems from data became popular in the last few years as it enables its users to accomplish several important tasks including interpretation, control and forecasting; see, for example, the monographs (Brunton et al., 2022; Kutz et al., 2016) for an introduction to these topics. The very same framework can also be used to forecast state *distributions* by means of the duality relation connecting the Koopman operator (which evolves states and observables) and the Perron-Frobenius operator, which evolves distributions. Kernel based algorithms to learn the Koopman or transfer operators have been studied in (Alexander & Giannakis, 2020; Bouvrie & Hamzi, 2017; Das & Giannakis, 2020; Klus et al., 2019; Kostic et al., 2022; 2023a; Meanti et al., 2023; Nüske et al., 2023; Williams et al., 2015; Bevanda et al., 2023; Hou et al., 2023). Deep learning approaches, on the other hand, were explored in the works (Bevanda et al., 2021; Fan et al., 2021; Kostic et al., 2023b; Lusch et al., 2018; Azencot et al., 2020; Morton et al., 2018).

Our aim it to improve over mainstream KOR estimators, whose performance is known to deteriorate as the forecasting horizon extends further into the future (Kostic et al., 2022). Within the setting of *uniquely ergodic* dynamical systems, arbitrary initial state distributions are bound to

[1]CMAP, Ecole Polytechnique, Palaiseau, France [2]Italian Institute of Technology, Genoa, Italy [3]University of Novi Sad, Serbia [4]University College London, UK. Correspondence to: Vladimir Kostic <vladimir.kostic@iit.it>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

[1]Historically, the Koopman operator was introduced for deterministic dynamical systems, while the transfer operator is its analogue in the stochastic case. The results presented in this paper, however, apply to both settings.

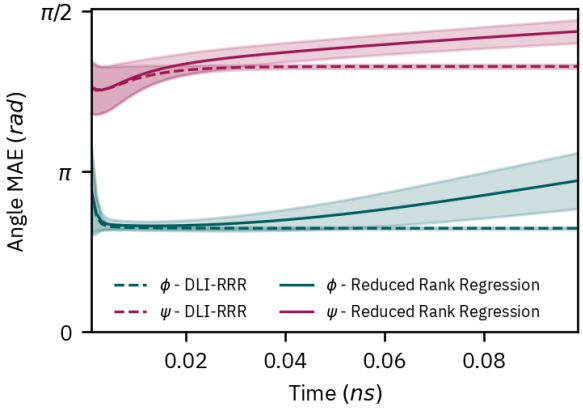

Figure 1. Mean Absolute Error (MAE) in forecasting the backbone dihedral angles of Alanine Dipeptide. Data points are 10^{-3} ns apart.

converge to an unknown but unique *invariant*[2] distribution. In this work we propose a paradigm to inject this prior knowledge into existing kernel-based algorithms, producing estimators which accurately forecast the long-term behavior of observables of the system. Figure 1 illustrates the benefit of such estimator on a molecular dynamics simulations experiment, in which we wish to forecast the dihedral angles of the Alanine Dipeptide (Wehmeyer & Noé, 2018) molecule over a long time horizon. The figure shows that the forecasting error of a state-of-the-art operator regression estimator deteriorates as the forecasting horizon increases. In contrast, when the same estimator is augmented by our *deflate-learn-inflate* (DLI) method, the forecasting error remains uniformly bounded in time, as predicted by our theoretical analysis.

Contributions Our conceptually simple approach builds upon well-known ideas in the literature (deflating and centering) and can be seamlessly integrated into any KOR estimator based on empirical risk minimization to enhance their long-term forecasting accuracy. Yet our principal contribution is to derive the first non-asymptotic forecasting error bounds that hold *uniformly over the time horizons*. We address both, forecasting the conditional mean of any observable, and forecasting the state distributions from an initial one.

Paper Organization Sec. 2 briefly reviews Koopman/ transfer operators and their estimators. Sec. 3 introduces the long-term forecasting problem alongside key quantities used to characterized the error induced by the estimators studied in the paper. In Sec. 4, we present the DLI approach and discuss its implementation, while Sec. 5 contains our theoretical guarantees. Sec. 6 addresses long-term distribution forecasting. Finally, in Sec. 7 we present numerical experiments with our approach.

[2]That is, invariant under the action of the Perron-Frobenius operator.

Notations If $\mathcal{H}$ is a separable Hilbert space, and $(e_i)_{i\in\mathbb{N}}$ an orthonormal basis, we let $\mathrm{HS}(\mathcal{H})$ be the Hilbert space of Hilbert-Schmidt (HS) operators on $\mathcal{H}$ endowed with the norm $\|A\|_{\mathrm{HS}}^2 \equiv \sum_{i\in\mathbb{N}} \|Ae_i\|_{\mathcal{H}}^2$, for $A \in \mathrm{HS}(\mathcal{H})$. For any bounded operator A on $\mathcal{H}$, we denote by $\rho(A)$ and $\|A\|$ the spectral radius and operator norm of A respectively. Note that $\rho(A) \leq \|A\|$ (see e.g. Trefethen & Embree, 2020). Finally, for two measures μ and ν, $\mu \ll \nu$ means that μ is absolutely continuous w.r.t. ν, in which case $d\mu/d\nu$ denotes the Radon-Nikodym derivative.

2. Background

In this section, we give some background on the transfer operators (Lasota & Mackey, 1994) and their empirical estimators (Kostic et al., 2022). Throughout the paper we study discrete stochastic dynamical systems, $(X_t)_{t\in\mathbb{N}}$, where the state at time $t \in \mathbb{N}$ forms a random variable X_t with law μ_t, taking values in a measurable space $\mathcal{X}$, endowed with σ-algebra $\Sigma_{\mathcal{X}}$. We assume that the sequence $(X_t)_{t\in\mathbb{N}}$ is a time-homogeneous Markov process, that is $\mathbb{P}[X_{t+1} \,|\, (X_s)_{s=0}^t] = \mathbb{P}[X_{t+1} \,|\, X_t]$, and there exists a *transition kernel* $p\colon \mathcal{X} \times \Sigma_{\mathcal{X}} \to [0,1]$, such that, for every $(x, B) \in \mathcal{X} \times \Sigma_{\mathcal{X}}$ and $t \in \mathbb{N}$,

$$\mathbb{P}[X_{t+1} \in B \,|\, X_t = x] = p(x, B).$$

We further assume that the dynamical system is *uniquely ergodic*, that is, there exists a *unique* probability distribution π, called *invariant measure*, such that if $X_0 \sim \pi$, then $X_t \sim \pi$, for every $t \in \mathbb{N}$.

Koopman Operator The above dynamical systems are general enough to capture several important phenomena, including (discretized) Langevin dynamics (Davidchack et al., 2015) or other systems constructed from the discretization of stochastic differential equations. They can be studied via Markov operators, and, in particular with *forward transfer operators* $A_\pi\colon L^2_\pi(\mathcal{X}) \to L^2_\pi(\mathcal{X})$ defined on the space $L^2_\pi(\mathcal{X})$ formed by square integrable functions w.r.t. the invariant measure as

$$[A_\pi f](x) := \mathbb{E}[f(X_{t+1}) \,|\, X_t = x],\ x \in \mathcal{X}, t \in \mathbb{N}. \quad (1)$$

Due to their prominence in the data-driven (deterministic) dynamical systems community (see e.g. Brunton et al., 2022), we also call A_π the (stochastic) Koopman operators.

The significance of Koopman operators lies in their ability to effectively *linearize* the underlying Markov processes. Namely, for every observable $f \in L^2_\pi(\mathcal{X})$, computing its expected value after t time steps from some initial state $x \in \mathcal{X}$ is simply powering of Koopman operator A_π, i.e.

$$\mathbb{E}[f(X_t) \,|\, X_0 = x] = [A_\pi^t f](x). \quad (2)$$

Perron-Frobenious Operator Additional interest in transfer operators comes from the duality between observables and state distributions. Specifically, if μ_0 is absolutely continuous w.r.t. the invariant measure π, that is, it has a density $q_0 := d\mu_0/d\pi \in L^1_\pi(\mathcal{X})$ defined via the Radon-Nikodyn derivative, and in addition the density is square-integrable, then, for every $t \in \mathbb{N}$ one has $q_t := d\mu_t/d\pi \in L^2_\pi(\mathcal{X})$, and the flow of the probability distributions $(q_t)_{t\in\mathbb{N}}$ follows *linear dynamics* in the space $L^2_\pi(\mathcal{X})$, given by the equations

$$q_t = A^*_t q_0 = (A^*_\pi)^t q_0, \quad t \in \mathbb{N}. \tag{3}$$

The operator A^*_π, known as *Perron-Frobenius operator*, is the adjoint of the Koopman operator, and it is given, for every $q \in L^2_\pi(\mathcal{X})$, by

$$[A^*_\pi q](y) = \int p^*(y, dx)q(x) \tag{4}$$

where p^* is the time-reversal transition kernel defined, for every $B \in \Sigma$ and $y \in \mathcal{X}$ as $p^*(y, B) = \mathbb{P}[X_{t-1} \in B | X_t = y]$. A consequence of (3) is that, once linearized, the process can be efficiently evolved from any initial density q_0 using the spectral theory of bounded operators.

Among all observables/densities, constant ones play a particular role. Namely, (1) and (4) imply that

$$A_\pi \mathbb{1}_\pi = \mathbb{1}_\pi, \quad \text{and} \quad A^*_\pi \mathbb{1}_\pi = \mathbb{1}_\pi, \tag{5}$$

where $\mathbb{1}_\pi \in L^2_\pi(\mathcal{X})$ is the function π-almost everywhere equal to 1. So, since $\|A_\pi\| = 1$, its largest eigenvalue is equal to 1, which since the process is uniquely ergodic, has (up to scaling) unique eigenfunction $\mathbb{1}_\pi$.

Operator Regression In recent years, the abundance of emerging machine learning algorithms has sparked a growing interest on data-driven dynamical systems. In this setting A_π is not known, and a key challenge is to learn it from data. An appealing class of KOR learning algorithms (Brunton et al., 2022; Kostic et al., 2022; Kutz et al., 2016) aim to learn the Koopman operator on a predefined reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ consisting of functions from $L^2_\pi(\mathcal{X})$. Namely, let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric and positive definite kernel function and $\mathcal{H}$ the corresponding RKHS (Aronszajn, 1950), with norm denoted as $\|\cdot\|_\mathcal{H}$. We let $x \mapsto k_x \equiv k(\cdot, x) \in \mathcal{H}$ denote the *canonical feature map* and assume, for every $x \in \mathcal{X}$, that $k(\cdot, x) \in L^2_\pi(\mathcal{X})$, which in turn implies that $\mathcal{H} \subset L^2_\pi(\mathcal{X})$; for an introduction to RKHS see, e.g., Steinwart & Christmann (2008).

In data-driven dynamical systems we are provided with a dataset $\mathcal{D}_n := (x_i, y_i)_{i=1}^n$ of consecutive states sampled at equilibrium, and wish to learn the Koopman operator by minimizing the *empirical risk*

$$\widehat{\mathcal{R}}(G) := \tfrac{1}{n} \textstyle\sum_{i\in[n]} \|k_{y_i} - G^* k_{x_i}\|^2, \tag{6}$$

over operators $G \colon \mathcal{H} \to \mathcal{H}$. Defining the sampling operators $\widehat{S}_x, \widehat{S}_y \colon \mathcal{H} \to \mathbb{R}^n$ as $\widehat{S}_x h := \frac{1}{\sqrt{n}}(h(x_i))_{i\in[n]}$ and $\widehat{S}_y h := \frac{1}{\sqrt{n}}(h(y_i))_{i\in[n]}$, (6) can be equivalently written as

$$\widehat{\mathcal{R}}(G) = \|\widehat{S}_y - \widehat{S}_x G\|^2_{\text{HS}}$$

from which it is apparent that the estimators are of the form $\widehat{G} = \widehat{S}^*_x W \widehat{S}_y$, for some $n \times n$ real matrix W. In particular, we mention three important estimators that are often used in applications: kernel ridge regression (KRR), principal component regression (PCR) and reduced rank regression (RRR), implementing different forms of regularization on the operator G. While they are defined via empirical covariance $\widehat{C} := \widehat{S}^*_x \widehat{S}_x$ and cross-covariance operators $\widehat{T} := \widehat{S}^*_x \widehat{S}_y$ on the RKHS $\mathcal{H}$, in practice, they are computed via kernel Gram matrices $K_x := \widehat{S}_x \widehat{S}^*_x = \frac{1}{n}[k(x_i,x_j)]_{i,j\in[n]}$ and $K_y := \widehat{S}_y \widehat{S}^*_y = \frac{1}{n}[k(y_i,y_j)]_{i,j\in[n]}$, and applied using $K_{xy} := \widehat{S}_x \widehat{S}^*_y = \frac{1}{n}[k(x_i,y_j)]_{i,j\in[n]}$, see (Kostic et al., 2022) for more information and the explicit form of estimators.

Estimation Error The spaces $\mathcal{H}$ and $L^2_\pi(\mathcal{X})$ have different norms. To handle this ambiguity, we use the *injection operator* $S_\pi : \mathcal{H} \hookrightarrow L^2_\pi(\mathcal{X})$ such that, for all $f \in \mathcal{H}$, the object $S_\pi f$ is the element of $L^2_\pi(\mathcal{X})$ which is pointwise equal to $f \in \mathcal{H}$, but endowed with the appropriate $L^2_\pi(\mathcal{X})$ norm. Moreover, the adjoint of the injection is given, for $f \in L^2_\pi(\mathcal{X})$, by

$$S^*_\pi f = \mathbb{E}_{X\sim\pi}[f(X)k_X] \in \mathcal{H}. \tag{7}$$

Every estimator $\widehat{G}$ defines an approximation of $A_{|\mathcal{H}} \colon \mathcal{H} \to L^2_\pi(\mathcal{X})$, the *restriction* of the Koopman operator to the RKHS, namely $A_{|\mathcal{H}} = A_\pi S_\pi$. Specifically, we estimate $A_{|\mathcal{H}}$ by the operator $S_\pi \widehat{G} \colon \mathcal{H} \to L^2_\pi(\mathcal{X})$. The corresponding estimation quality can be measured by the operator norm error

$$\mathcal{E}(\widehat{G}) := \|A_\pi S_\pi - S_\pi \widehat{G}\|, \tag{8}$$

that is upper bounded by the excess risk of the operator regression problem formulated in (Kostic et al., 2022). When $\mathcal{H}$ is an infinite-dimensional *universal* RKHS optimal rates in the error (8) for all the above estimators above were developed in Kostic et al. (2023a).

3. Long-term Forecasting

While Koopman operator estimators are learned to guarantee good one-step-ahead prediction, in order to *truly learn dynamics*, recalling (2), one needs to guarantee good *long-term forecasting*, that is $[A^t_\pi S_\pi h](x) \approx [\widehat{G}^t h](x)$, for $t \in \mathbb{N}$ and $h \in \mathcal{H}$, which we address in this section.

As stated in Kostic et al. (2022, Thm. 1), the control of the one-step-ahead error (8) is not enough to guarantee good

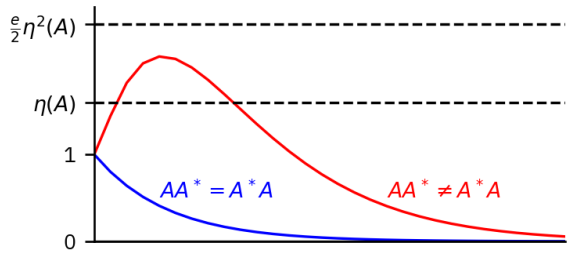

Figure 2. Asymptotically stable operators: For a a non-normal ($AA^* \neq A^*A$) asymptotically stable ($\rho(A) < 1$) operator A, $\|A^t\|$ exhibits transient growth captured by the Kreiss constant $\eta(A)$ before converging to zero at the linear rate $\rho(A)$.

forecasting for long time-horizons. Indeed,

$$\|\mathbb{E}[h(X_t) \,|\, X_0 = \cdot] - S_\pi \widehat{G}^t h\|_{L^2_\pi(\mathcal{X})} \leq \mathcal{E}_t(\widehat{G})\, \|h\|_{\mathcal{H}},$$

where $\mathcal{E}_t(\widehat{G}) := \|A^t_\pi S_\pi - S_\pi \widehat{G}^t\|$, $t \in \mathbb{N}$ is the error induced by the *power* of the estimator.

After some algebra one verifies that $A^t_\pi S_\pi - S_\pi \widehat{G}^t = \sum_{k=0}^{t-1} A^k_\pi (A_\pi S_\pi - S_\pi \widehat{G}) \widehat{G}^{t-1-k}$. Hence applying the norm we have that

$$\mathcal{E}_t(\widehat{G}) \leq \min\{s(A_\pi)\, p(\widehat{G}), s(\widehat{G})\, p(A_\pi)\}\, \mathcal{E}(\widehat{G}), \quad (9)$$

where for an operator A we define

$$s(A) := \sum_{t=0}^{\infty} \|A^t\| \quad \text{and} \quad p(A) := \sup_{t \in \mathbb{N}_0} \|A^t\|. \quad (10)$$

Therefore, to obtain long-term consistent forecasting, apart from bounding the error (8), the two quantities in (10) also need to be bounded. Unfortunately, whenever $\rho(A) = 1$ we have that $s(A) = +\infty$. Hence, recalling (5), for a consistent estimator $\widehat{G}$ of A_π we have that $s(A_\pi) = \infty$ and $s(\widehat{G}) \to \infty$ with the number of samples, presenting a difficulty in obtaining bounds on the infinite time-horizons. Moreover, whenever the leading eigenvalue is not perfectly estimated as 1, long term forecasting either explodes ($\rho(\widehat{G}) > 1$) or collapses to zero ($\rho(\widehat{G}) < 1$). In order to overcome this issue, we resort to well-known concepts in the study of asymptotically stable linear dynamical systems in a Hilbert space, see e.g. (Trefethen & Embree, 2020).

For a bounded linear operator A such that $\rho(A) < 1$ we have that $\lim_{t\to\infty} \|A^t\| = 0$, but, depending on the *normality* of the operator, the convergence might not be monotone. Namely, if A is a normal operator, that is $AA^* = A^*A$, then $\|A^t\| = [\rho(A)]^t$, as illustrated by the blue line in Figure 2, and consequently, $p(A) = 1$ and $s(A) = 1/(1-\rho(A))$. Moreover, in this case $1/s(A)$ coincides with the *distance to instability* of A

$$d(A) := \inf_{z \in \mathbb{C},\, |z| \geq 1} \|(A - zI)^{-1}\|^{-1} \quad (11)$$

that measures the distance of the operator's spectra to the unit circle relative to its sensitivity to perturbations, which for normal operators equals $1-\rho(A)$.

On the other hand, as illustrated by the red line in Fig. 2, when A is a non-normal operator, the sequence $(\|A^t\|)_{t\in\mathbb{N}_0}$ may exhibit a *transient growth* before converging to zero, that can be estimated by $\eta(A) \leq p(A) \leq (e/2)[\eta(A)]^2$ (El-Fallah & Ransford, 2002), where $\eta(A)$ is the *the Kreiss* constant of A defined as

$$\eta(A) := \sup_{z \in \mathbb{C},\, |z|>1} (|z|-1)\|(A - zI)^{-1}\| \geq 1. \quad (12)$$

For highly non-normal operators $\eta(A) \gg 1$, indicating a large transient growth, which is also related to much smaller distance to instability $d(A) \ll 1 - \rho(A)$, and larger cumulative effect $s(A) \gg 1/(1-\rho(A))$. Nevertheless, the latter quantity always remains bounded, since due to $\limsup_{t\to\infty} \|A^t\|^{1/t} = \rho(A) < 1$, there exists the smallest integer ℓ such that $\|A^\ell\| < 1$, and, consequently, $s(A) \leq \frac{1}{1-\|A^\ell\|} \frac{\|A\|^\ell - 1}{\|A\| - 1} < \infty$.

Therefore, a promising approach to derive forecasting bounds *independent of the time-horizon* is to transform the learning objective from nonexpansive ($\|A\| = 1$) to asymptotically stable ($\rho(A) < 1$), which we introduce in the following section.

4. Deflate-Learn-Inflate (DLI) Estimators

In section we present a conceptually simple *estimator agnostic method* which overcomes the long term forecasting failure of Koopman operator regression (KOR). It consists of three steps: i) Remove the leading eigenvalue from the transfer operator (deflate); ii) Compute an estimator from data using centered features (learn); iii) Evolve the observable with such estimator and correct it using the averages over training data-points (inflate). We proceed to explain each of these steps in turn.

Deflate The first step is a classical idea in the field of numerical methods for eigenvalue problems, see, e.g., (Saad, 2011). In our context, recalling (5), it consists of removing (*deflating*) the known eigenpair $(1, \mathbb{1}_\pi)$ of the Koopman operator A_π, in order to better estimate the unknown ones. Since the leading Koopman eigenvalue $\lambda_1(A_\pi) = 1$ is simple, its corresponding spectral projector is $\mathbb{1}_\pi \otimes \mathbb{1}_\pi$. The corresponding *deflated* operator is

$$\mathbf{A}_\pi := A_\pi - \mathbb{1}_\pi \otimes \mathbb{1}_\pi = A_\pi J_\pi = J_\pi A_\pi, \quad (13)$$

where $J_\pi := I - \mathbb{1}_\pi \otimes \mathbb{1}_\pi$ is the orthogonal projector onto the orthogonal complement of the subspace of constant functions. Then, for every $t \in \mathbb{N}$, $\mathbf{A}^t_\pi = A^t_\pi - \mathbb{1}_\pi \otimes \mathbb{1}_\pi$ implies

$$\mathbb{E}[h(X_t) \,|\, X_0 = \cdot] = \mathbf{A}^t_\pi S_\pi h + \mathbb{E}_{X\sim\pi}[h(X)]. \quad (14)$$

Learn To learn the deflated operator we follow the same RKHS approach of operator regression, that is we approximate $\mathbf{A}_{\pi|_{\mathcal{H}}} = \mathbf{A}_\pi S_\pi = J_\pi A_\pi J_\pi S_\pi$. Noting that $\mathrm{Im}(\mathbf{A}_\pi) \subseteq J_\pi$ we define the "injection" to the appropriate subspace $\mathbf{S}_\pi \colon \mathcal{H} \to \mathrm{Im}(J_\pi)$ as inject-and-project $\mathbf{S}_\pi := J_\pi S_\pi$, and look for the estimator $\widehat{\mathbf{G}} \colon \mathcal{H} \to \mathcal{H}$ that minimizes the estimation error of the deflated operator as

$$\mathcal{E}^\circ(\widehat{\mathbf{G}}) := \|\mathbf{A}_\pi \mathbf{S}_\pi - \mathbf{S}_\pi \widehat{\mathbf{G}}\|. \tag{15}$$

But then, the estimation error is controlled by the corresponding risk minimization, see Appendix A.4 for detailed derivation, where the empirical risk functional is given by

$$\widehat{\mathcal{R}}^\circ(\widehat{\mathbf{G}}) := \|\widehat{\mathbf{S}}_y - \widehat{\mathbf{S}}_x \widehat{\mathbf{G}}\|^2_{\mathrm{HS}}, \tag{16}$$

via the projected sampling operators are $\widehat{\mathbf{S}}_x = J_n \widehat{S}_x$ and $\widehat{\mathbf{S}}_y = J_n \widehat{S}_y$, matrix J_n being the orthogonal projection $J_n = I - \mathbb{1}_n \mathbb{1}_n^\top$, and the vector $\mathbb{1}_n = n^{-1/2}[1, 1 \ldots, 1]^\top \in \mathbb{R}^n$.

Next, observing that for the i-th standard basis vector $e_i \in \mathbb{R}^n$, we have that

$$\widehat{\mathbf{S}}_x^* e_i = \widehat{S}_x^*(e_i - \tfrac{1}{\sqrt{n}}\mathbb{1}_n) = \tfrac{1}{\sqrt{n}}(k_{x_i} - \tfrac{1}{n}\textstyle\sum_{j\in[n]} k_{x_j})$$

and similarly that $\widehat{\mathbf{S}}_y^* e_i = \frac{1}{\sqrt{n}}(k_{y_i} - \frac{1}{n}\sum_{j\in[n]} k_{y_j})$. Using these, we can rewrite the empirical risk (16) as

$$\widehat{\mathcal{R}}^\circ(\widehat{\mathbf{G}}) := \tfrac{1}{n}\sum_{i\in[n]} \left\| \left(k_{y_i} - \tfrac{1}{n}\sum_{j\in[n]} k_{y_j}\right) - \widehat{\mathbf{G}}^* \left(k_{x_i} - \tfrac{1}{n}\sum_{j\in[n]} k_{x_j}\right) \right\|^2$$

which shows an elegant connection between the two learning problems. Namely, any estimator $\widehat{G}$ of the Koopman operator A_π can be transformed into an estimator $\widehat{\mathbf{G}}$ of the deflated Koopman operator $\mathbf{A}_\pi$ by simply *empirically centering the feature map* of the kernel. In practice, this means instead of using kernel Gram matrices K_x and K_y, to use their centered versions $\mathbf{K}_x := J_n K_x J_n$ and $\mathbf{K}_y := J_n K_y J_n$, respectively. Moreover, as we show in Section 5, the deflated version $\widehat{\mathbf{G}}$ of an estimator $\widehat{G}$ can readily be statistically studied by replacing Koopman operator A_π, injection S_π and sampling operators $\widehat{S}_x$ and $\widehat{S}_y$ by the projected ones $\mathbf{A}_\pi$, $\mathbf{S}_\pi$, $\widehat{\mathbf{S}}_x$ and $\widehat{\mathbf{S}}_y$, respectively.

Inflate Finally, we use an estimator $\widehat{\mathbf{G}}$ to obtain the empirical estimates of any observable $h \in \mathcal{H}$ as $\mathbb{E}[h(X_t) \,|\, X_0 = \cdot]$. For this purpose, we need to put back (*inflate*) the leading eigenpair that we removed during the deflate step. Namely, recalling (14), since $\mathbf{A}_\pi S_\pi h \approx J_\pi S_\pi \widehat{\mathbf{G}}^t h = S_\pi \widehat{\mathbf{G}}^t h - \langle \mathbb{1}_\pi, S_\pi \widehat{\mathbf{G}}^t h \rangle \mathbb{1}_\pi$ and $\mathbb{E}_{X\sim\pi}[h(X)] = \langle S_\pi^* \mathbb{1}_\pi, h\rangle_{\mathcal{H}}$, we have

$$A_\pi S_\pi h \approx S_\pi \widehat{\mathbf{G}}^t h + \langle S_\pi^* \mathbb{1}_\pi, h - \widehat{\mathbf{G}}^t h\rangle_{\mathcal{H}} \mathbb{1}_\pi. \tag{17}$$

Thus, we have an additional term in estimation that depends on (unknown) invariant measure. However, recalling (7), $S_\pi^* \mathbb{1}_\pi$ is equal to

$$k_\pi := \int_{\mathcal{X}} k_x \pi(dx) = \mathbb{E}_{X\sim\pi}[k_X] \in \mathcal{H}, \tag{18}$$

known as the *kernel mean embedding* (KME) of the measure π for which we have empirical estimators

$$\hat{\pi}_x := \tfrac{1}{n}\textstyle\sum_{i\in[n]} \delta_{x_i} \quad \text{and} \quad \hat{\pi}_y := \tfrac{1}{n}\sum_{i\in[n]} \delta_{y_i}, \tag{19}$$

associated to the input and the output points, respectively. Thus, we can approximate k_π with both $k_{\hat{\pi}_x}$ and $k_{\hat{\pi}_y}$ to estimate

$$\langle k_\pi, h - \widehat{\mathbf{G}}^t h\rangle_{\mathcal{H}} \approx \langle k_{\hat{\pi}_y}, h\rangle_{\mathcal{H}} - \langle k_{\hat{\pi}_y}, \widehat{\mathbf{G}}^t h\rangle_{\mathcal{H}}, \tag{20}$$

and, consequently, estimate $\mathbb{E}[h(X_t) \,|\, X_0 = x]$ by

$$\widehat{h}_t(x) := [\widehat{\mathbf{G}}^t h](x) + \tfrac{1}{n}\sum_{i\in[n]} \Big(h(y_i) - [\widehat{\mathbf{G}}^t h](x_i)\Big), x \in \mathcal{X}. \tag{21}$$

Since we always have estimators of finite rank $r \leq n$ obtained by minimizing the empirical risk $\widehat{\mathcal{R}}^\circ$, they are of the form $\widehat{\mathbf{G}} = \widehat{\mathbf{S}}_x^* U_r V_r^\top \widehat{\mathbf{S}}_y$ for some $U_r, V_r \in \mathbb{R}^{n\times r}$. Thus, after some algebra, we obtain that powering of the estimator can be efficiently computed by powering $r \times r$ matrix $M := U_r^\top \mathbf{K}_{xy} V_r \in \mathbb{R}^{r\times r}$ to obtain

$$\widehat{h}_t(x) = \sum_{i,j\in[n]} [w_{t,j} + (W_t)_{ij} k(x_j, x)] h(y_i),\ x \in \mathcal{X},$$

where the weights matrix is computed as

$$W_t := J_n V_r M^{t-1} (J_n U_r)^\top \in \mathbb{R}^{n\times n} \tag{22}$$

and vector as

$$w_t := \tfrac{1}{\sqrt{n}}(\mathbb{1}_n - W_t K_x \mathbb{1}_n) \in \mathbb{R}^n. \tag{23}$$

Note that the computational complexity added by DLI is modest as it only amounts to centering the kernels with a cost of $O(n^2)$, while the steps (22) and (23) incur the same computational complexity as for the standard estimator ($J_n = I$). Appendix A.4 gives a detailed algorithm for DLI versions of KRR, PCR and RRR estimators.

5. Time-independent forecasting bounds

In this section we prove that the DLI paradigm transforms Koopman estimators which are only one step ahead consistent, into estimators that achieve uniform consistency over time according to either $L^2_\pi(\mathcal{X})$-error for the observables.

First, since $\rho(\mathbf{A}_\pi) < 1$, and hence $s(\mathbf{A}_\pi) < \infty$, the following result indicates that statistical bounds for one-step-ahead operator norm error $\mathcal{E}^\circ$ and supremum p of the estimator are sufficient for uniformly good forecasting.

Proposition 5.1. *Let $\widehat{\boldsymbol{G}}$ be any estimator of $\boldsymbol{A}_\pi$, and let $\widehat{h}_t$ be given by (21). Then, for every $h \in \mathcal{H}$ and $t \in \mathbb{N}$*

$$\|A_\pi^t S_\pi h - S_\pi \widehat{h}_t\|/\|h\| \leq p(\widehat{\boldsymbol{G}}) \left(s(\boldsymbol{A}_\pi)\, \mathcal{E}^\circ(\widehat{\boldsymbol{G}}) + \varepsilon \right) + \varepsilon,$$

where $\max\{\|k_\pi - k_{\hat{\pi}_x}\|, \|k_\pi - k_{\hat{\pi}_y}\|\} \leq \varepsilon$ for some $\varepsilon > 0$.

Proof. Observe that (17) and (21) imply, by adding and subtracting $k_{\hat{\pi}_x}$ and $k_{\hat{\pi}_y}$, that the forecasting error $\|A_\pi^t S_\pi h - S_\pi \widehat{\mathbf{G}}^t h\|/\|h\|$ can be upper bounded by $\|\mathbf{A}_\pi^t \mathbf{S}_\pi - \mathbf{S}_\pi \widehat{\mathbf{G}}^t\| + \|\widehat{\mathbf{G}}^t\| \|k_\pi - k_{\hat{\pi}_x}\| + \|k_\pi - k_{\hat{\pi}_y}\|$. Hence, (9) applied to $\mathcal{E}_t^\circ$ completes the proof. □

Next, we state our main assumptions on the centered operators. We use the symbol "$*$" to distinguish assumptions on the centered operators $\mathbf{C}, \mathbf{T}$ from their standard (uncentered) operators C and T.

- **(BK)** *Boundedness.* There exists $c_{\mathcal{H}} > 0$ such that $\operatorname{ess\,sup}_{x \sim \pi} k(x,x) \leq c_{\mathcal{H}}$.
- **(RC*)** *Regularity Condition.* For some $\alpha \in [1,2]$ there exists $a > 0$ such that $\mathbf{T}\mathbf{T}^* \preceq a^2 \mathbf{C}^{1+\alpha}$.
- **(SD*)** *Spectral Decay.* There exists $\beta \in (0,1]$ and a constant $b > 0$ so that $\lambda_j(\mathbf{C}) \leq b\, j^{-1/\beta}$, $\forall j \in \mathbb{N}$.

Assumptions **(BK)** and **(SD)** are taken from the works (Fischer & Steinwart, 2020; Li et al., 2022) on kernel mean embeddings. Assumption **(RC)** was introduced in (Kostic et al., 2023a) to study KOR estimators.

First, note that $\mathbf{C} = \mathbf{S}_\pi^* \mathbf{S}_\pi = S_\pi^* J_\pi S_\pi \preceq S_\pi^* S_\pi = C$, and $\operatorname{Im}(A_\pi S_\pi) \subseteq \operatorname{Im}(S_\pi)$ implies $\operatorname{Im}(\mathbf{A}_\pi \mathbf{S}_\pi) \subseteq \operatorname{Im}(\mathbf{S}_\pi)$. Thus, we have that conditions **(RC*)** and **(SD*)** are weaker than conditions **(RC)** and **(SD)**, respectively. That is, if C and T satisfy **(RC)** and **(SD)**, then the centered objects $\mathbf{C}$ and $\mathbf{T}$ satisfy **(RC*)** and **(SD*)** with possibly smaller β.

In this paper we provide analysis for the centered KRR, PCR and RRR estimators. Since in the DLI method, we only need to replace $\widehat{C}$ and $\widehat{T}$ by their centered version $\widehat{\mathbf{C}}$ and $\widehat{\mathbf{T}}$ respectively to define the deflated versions $\widehat{\mathbf{G}}_\gamma, \widehat{\mathbf{G}}_{r,\gamma}^{\mathrm{PCR}}, \widehat{\mathbf{G}}_{r,\gamma}^{\mathrm{RRR}}$, in order to optimally control the error (15) we rely on the proof techniques of Kostic et al. (2023a), based on the SVD decomposition of the injection operator S_π. Since the main difference between centered and uncentered operators arises from using the projected injection operator $\mathbf{S}_\pi$ instead of S_π, their analysis can be extended in an elegant way to control $\mathcal{E}^\circ(\widehat{\mathbf{G}})$. To this end, we only need to extend two bounds on whitened features to the centered case.

Proposition 5.2. *Let **(BK)** and **(SD*)** hold for some $\beta \in (0,1]$, and let $\xi(x) := \boldsymbol{C}_\gamma^{-1/2}[k_x - k_\pi]$. Then there exist $\tau \in [\beta, 1]$ and $c_\tau, c_\beta \in (0,\infty)$ such that for $\gamma > 0$*

$$\|\xi\|^2 \leq c_\beta \gamma^{-\beta} \quad \text{and} \quad \|\xi\|_\infty^2 \leq c_\tau \gamma^{-\tau}. \tag{24}$$

Therefore, variance control of estimators from (Kostic et al., 2023a) can be readily applied which leads to the learning rates for centered KRR, RRR and PCR estimators.

To address the more challenging problem of bounding the transient growth of the empirical estimators, we use the following result, showing how one can obtain the concentration of the estimators' Kreiss constant in the RKHS operator norm; see App. B for the proof.

Lemma 5.3. *Let **(RC*)** hold for some $\alpha > 1$, then there exists a compact $\boldsymbol{G}_{\mathcal{H}}: \mathcal{H} \to \mathcal{H}$ such that $\boldsymbol{A}_\pi \boldsymbol{S}_\pi = \boldsymbol{S}_\pi \boldsymbol{G}_{\mathcal{H}}$ and $\rho(\boldsymbol{G}_{\mathcal{H}}) < 1$. Consequently, $\eta(\boldsymbol{G}_{\mathcal{H}}) < \infty$, $d(\boldsymbol{G}_{\mathcal{H}}) > 0$ and*

$$\frac{\eta(\boldsymbol{G}_{\mathcal{H}}) d(\boldsymbol{G}_{\mathcal{H}})}{d(\boldsymbol{G}_{\mathcal{H}}) + \|\boldsymbol{G}_{\mathcal{H}} - \widehat{\boldsymbol{G}}\|} \leq p(\widehat{\boldsymbol{G}}) \leq \frac{e}{2} \left[\frac{\eta(\boldsymbol{G}_{\mathcal{H}}) d(\boldsymbol{G}_{\mathcal{H}})}{d(\boldsymbol{G}_{\mathcal{H}}) - \|\boldsymbol{G}_{\mathcal{H}} - \widehat{\boldsymbol{G}}\|} \right]^2.$$

holds for every $\widehat{\boldsymbol{G}}$ such that $\|\boldsymbol{G}_{\mathcal{H}} - \widehat{\boldsymbol{G}}\| < d(\boldsymbol{G}_{\mathcal{H}})$.

We now specify Thm. 5.1 for the KRR estimator, i.e. when $U_r = I$ and $V_r = (\mathbf{K}_x + \gamma I)^{-1}$, for some regularization parameter $\gamma > 0$, and iid samples from the invariant distribution. The results holding for the PCR and RRR estimators with $\alpha \in [1,2]$, as well as *realistic non-iid sampling along a trajectory* of a beta-mixing process, are given in Appendix B.

Theorem 5.4. *Let **(SD*)** and **(RC*)** hold for some $\beta \in (0,1]$ and $\alpha \in (1,2]$, respectively. In addition, let $\operatorname{cl}(\operatorname{Im}(S_\pi)) = L_\pi^2(\mathcal{X})$ and **(BK)** be satisfied. If $\delta \in (0,1)$,*

$$\gamma \asymp n^{-\frac{1}{\alpha+\beta}} \quad \text{and} \quad \varepsilon_n^\star := n^{-\frac{\alpha}{2(\alpha+\beta)}}, \tag{25}$$

then, for every $t \in \mathbb{N}$, the forecasted observable given in (21) based on KRR satisfies

$$\|\mathbb{E}[h(X_t) \,|\, X_0 = \cdot] - S_\pi \widehat{h}_t\|/\|h\| \leq C \varepsilon_n^\star \ln(\delta^{-1}),$$

with probability at least $1 - \delta$ w.r.t. iid sampled data $\mathcal{D}$ according to the invariant distribution π, where the constant C may depend only on a, b and $c_{\mathcal{H}}$.

Proof Sketch. Recall that the centered population (**KRR**) model is defined as $\mathbf{G}_\gamma = \mathbf{C}_\gamma^{-1}\mathbf{T}$ where $\mathbf{C}_\gamma := \mathbf{C} + \gamma I_{\mathcal{H}}$, while the empirical estimator is $\widehat{\mathbf{G}}_\gamma = \widehat{\mathbf{C}}_\gamma^{-1}\widehat{\mathbf{T}}$. Now, in view of Thm. (5.1) and Lem. 5.3, it suffices to prove that $\|\mathbf{G}_{\mathcal{H}} - \widehat{\mathbf{G}}_\gamma\| \leq d(\mathbf{G}_{\mathcal{H}})/2$ w.h.p. and derive the learning rate for $\mathcal{E}^\circ(\widehat{\mathbf{G}}_\gamma)$. First, since, $\|\mathbf{G}_{\mathcal{H}} - \widehat{\mathbf{G}}_\gamma\| \leq \|\mathbf{G}_{\mathcal{H}} - \mathbf{G}_\gamma\| + \|\mathbf{G}_\gamma - \widehat{\mathbf{G}}_\gamma\|$, Lem. B.6 in App. B gives that $\|\mathbf{G}_{\mathcal{H}} - \mathbf{G}_\gamma\| \leq a\gamma^{(\alpha-1)/2}$. Next, using Kostic et al. (2023a, Proposition 16), for the KRR estimator we obtain that $\mathbb{P}\left\{ \|\widehat{\mathbf{G}}_\gamma - \mathbf{G}_\gamma\| \leq C \frac{\ln(2\delta^{-1})}{\sqrt{n\gamma^{\beta+1}}} \right\} \geq 1 - \delta$. Thus, provided that $\alpha > 1$ and n is taken large enough, with our choice of γ, we can guarantee that $\mathbb{P}\left\{ \|\mathbf{G}_{\mathcal{H}} - \widehat{\mathbf{G}}_\gamma\| \leq d(\mathbf{G}_{\mathcal{H}})/2 \right\} \geq 1 - \delta$.

Next, by applying Propositions 5 and 16 from Kostic et al. (2023a), we obtain that $\mathcal{E}^{\circ}(\widehat{\mathbf{G}}_{\gamma}) \leq C\varepsilon_n^{\star} \ln(\delta^{-1})$. Finally, Briol et al. (2019, Lem. 1) guarantees with probability at least $1-\delta$ that $\max\{\|k_{\pi} - k_{\hat{\pi}_x}\|, \|k_{\pi} - k_{\hat{\pi}_y}\|\} \leq \epsilon$ with $\epsilon = \sqrt{\frac{2}{n} c_{\mathcal{H}}} \left(1 + \sqrt{\log(\delta^{-1})}\right)$, and, hence, a union bound combining the previous results with Thm. 5.1 yields the result with probability at least $1-4\delta$. Up to a rescaling of the constants, we can replace $1-4\delta$ by $1-\delta$. □

We conclude this section noting that the previous theorem on the $L^2_{\pi}(\mathcal{X})$ estimation error of the conditional mean can be easily converted to the result on the $L^1_{\pi}(\mathcal{X})$ estimation error of the conditional variance. Namely, applying the estimator (21) to the squared observable we can estimate the second moment and easily derive the approximation of $\mathbb{V}[h(X_t)|X_0 = \cdot] := \mathbb{E}[h(X_t)^2|X_0 = \cdot] - (\mathbb{E}[h(X_t)|X_0 = \cdot])^2$, as demonstrated in Appendix B.6.

6. State Distribution Forecasting

In this section we show how DLI estimators of the Koopman operator defined on a universal RKHS can reliably be used as estimators of a Perron-Frobenious operator.

The RKHS framework naturally allows one to introduce a metric on the space of signed measures $\mathcal{M}^+(\mathcal{X})$ via kernel mean embeddings (see e.g. Muandet et al., 2017). That is, given an $\mathcal{H}$ we can define the dual norm

$$\|\mu\|_{\mathcal{H}^*} := \sup_{\|h\|_{\mathcal{H}} \leq 1} \int_{\mathcal{X}} h(x)\mu(dx), \ \mu \in \mathcal{M}^+(\mathcal{X}), \quad (26)$$

that induces the weak* topology on $\mathcal{M}^+(\mathcal{X})$. Recalling (18), the above supremum is attained at the KME k_{μ} of the measure μ. Moreover, the square distance of the kernel mean embeddings of two signed measures μ and ν,

$$\|\mu - \nu\|^2_{\mathcal{H}^*} = \|k_{\mu} - k_{\nu}\|^2_{\mathcal{H}}, \ \mu, \nu \in \mathcal{M}^+(\mathcal{X}), \quad (27)$$

is called the *maximum mean discrepancy* (MMD).

Next, we present how DLI estimators solve the problem of state distribution forecasting. Recalling that $q_t = d\mu_t/d\pi$, we have that $\mathbb{E}[h(X_t)] = \langle q_t, S_{\pi} h\rangle$ for $h \in \mathcal{H}$ and $t \in \mathbb{N}$. So, using (7) we have that $k_{\mu_t} = S^*_{\pi} q_t$, and, hence

$$\langle k_{\mu_t}, h\rangle = \mathbb{E}_{x\sim\mu_0}[\mathbb{E}[h(X_t)|X_0 = x]] = \langle q_0, A^t_{\pi} S_{\pi} h\rangle \quad (28)$$

$$\approx \langle q_0, S_{\pi} \widehat{h}_t\rangle = \langle k_{\mu_0} - k_{\hat{\pi}_x}, \widehat{\mathbf{G}}^t h\rangle + \langle k_{\hat{\pi}_y}, h\rangle_{\mathcal{H}} \quad (29)$$

$$\approx \langle k_{\widehat{\mu}_0} - k_{\hat{\pi}_x}, \widehat{\mathbf{G}}^t h\rangle + \langle k_{\hat{\pi}_y}, h\rangle_{\mathcal{H}} = \langle k_{\widehat{\mu}_t}, h\rangle \quad (30)$$

where the first approximation is according to Proposition 5.1, the second one uses an empirical estimate $\widehat{\mu}_0 = n_0^{-1} \sum_{i\in[n_0]} \delta_{z_i}$ of the initial measure μ_0, and the estimated KME is defined as

$$k_{\widehat{\mu}_t} := k_{\hat{\pi}_y} + (\widehat{\mathbf{G}}^*)^t (k_{\widehat{\mu}_0} - k_{\hat{\pi}_x}). \quad (31)$$

In this way, recalling (22)-(23), we have obtained $\mu_t \approx \widehat{\mu}_t := \sum_{j\in[n]} m_{t,j} \delta_{y_j}$, where the weights vector is $m_t = w_t + w^0_t \in \mathbb{R}^n$, for $w^0_t := \frac{1}{\sqrt{n}} W_t K_{xz} \mathbb{1}_{n_0}$ computed using the kernel Gram matrix $K_{xz} := \frac{1}{\sqrt{n_0 n}} [k(x_i, z_j)]_{i\in[n], j\in[n_0]}$.

A direct consequence of this construction is that, by taking the supremum of (28)-(30) over $h \in \mathcal{H}$, we can easily adapt Theorem 5.4 to bounding the MMD error between distributions as follows.

Theorem 6.1. *Under the assumptions of Theorem 5.4 for every $q_0 \in L^2_{\pi}(\mathcal{X})$ and $t \in \mathbb{N}$, with probability at least $1-\delta$ w.r.t. iid samples $\mathcal{D}_n$ according to π and samples $(z_i)_{i\in[n_0]}$ from the initial distribution μ_0, it holds*

$$\|\widehat{\mu}_t - \mu_t\|_{\mathcal{H}^*} \leq C \left(\frac{\ln(\delta^{-1})}{n^{\frac{\alpha}{2(\alpha+\beta)}}} + \sqrt{\frac{\ln \delta^{-1}}{n_0 \wedge n}} \right),$$

where the constant C may depend only on $a, b, c_{\mathcal{H}}$ and $\|q_0\|$.

According to Thm. 6.1, the DLI paradigm enables learning operators that can reliably forecast future state distributions, uniformly over time. Notice that in practice, we can easily sample from μ_0, so we can make $n_0 \geq n$ and then the dominating term in (6.1) is $\varepsilon^{\star}_n$, which depends only on the properties of kernel embedding and transfer operator.

Finally, an outstanding property of DLI estimators of state distributions is the *preservation of the probability mass* along all trajectory, since it holds that $\sum_{j\in[n]} m_{t,j} = 1$ due to the properties of the projector J_n.

7. Experiments

In this section, we compare the standard RRR estimator to its DLI-enhanced counterpart. Our results indicate that the DLI paradigm boosts the performance of the bare RRR model in long-term forecasting across the board.

CIR Model The Cox–Ingersoll–Ross (Cox et al., 1985) model (CIR) pertains to the field of mathematical finance and is routinely used to describe the evolution of interest rates. The CIR model characterizes the instantaneous interest rate r_t through the stochastic differential equation $dr_t = a(b - r_t)dt + \sigma\sqrt{r_t} dW_t$, where W_t is a Wiener process. In the CIR model, the interest rate adjusts to the mean b with a speed a. The volatility is described by σ and by the value of the rate itself through the term $\sqrt{r_t}$. For the CIR model, the conditional expectation of the state $\mathbb{E}[r_t \mid r_0 = \cdot]$ and its variance $\mathbb{V}[r_t \mid r_0 = \cdot]$ are known analytically (see Appendix). In Table 1 we report the root mean square error of RRR and DLI-RRR in estimating the conditional expectation and conditional variance. For both quantities, the DLI estimator attains smaller errors. Conditional mean and variance are estimated at $t = \ln 2/a$, corresponding to the half-life of the mean reversion, and are averaged over 100

Observable	RRR	**DLI-RRR**
$\mathbb{E}[r_t \mid r_0 = \cdot]$	0.0691 ± 0.0333	$\mathbf{0.0673 \pm 0.0328}$
$\mathbb{V}[r_t \mid r_0 = \cdot]$	0.0470 ± 0.0413	$\mathbf{0.0124 \pm 0.0051}$

Table 1. RMSE in estimating conditional expectation and variance of the CIR model (100 independent training datasets).

independent models trained on datasets of 500 points each. For this example we have set $a = 2.5$, $b = 1.0$ and $\sigma = 0.5$. To simulate the CIR model we discretized the stochastic differential equation with $\Delta t = 0.01$.

Ornstein-Uhlenbeck Model We study the uniformly sampled Ornstein-Uhlenbeck process $dX_t = -\theta X_t dt + \sigma dW_t$, where $\theta, \sigma > 0$ and W_t is a Wiener process. Integrating the stochastic differential equation shows that the probability flow of X_t (given that $X_0 = x$) is $\mu_t = \mathcal{N}\big(xe^{-\theta t}, \frac{\sigma^2}{2\theta}(1 - e^{-2\theta t})\big)$, yielding an invariant measure $\pi = \mathcal{N}(0, \sigma^2/2\theta)$. We investigate the equilibration of an Ornstein-Uhlenbeck process with initial condition X_0 drawn from a Gaussian mixture with means $\{-2, 2\}$ and variances $\{0.04, 0.04\}$, respectively. In Fig. 3 we report the predicted probability flow, as well as the *relative* MMD $\|\widehat{\mu}_t - \mu_t\|^2_{\mathcal{H}^*} / \|\mu_t\|^2_{\mathcal{H}^*}$ attained by DLI and uncentered estimators. Note that the DLI paradigm leads to a consistent improvement in forecasting performance throughout the entire trajectory. Notice how the MMD error exhibits a transient growth for both the RRR model and its DLI version during the short-term forecasting. Recalling Figure 2, this effect is present due to the non-normality of the estimators in the chosen RKHS space, in concordance with our theoretical analysis.

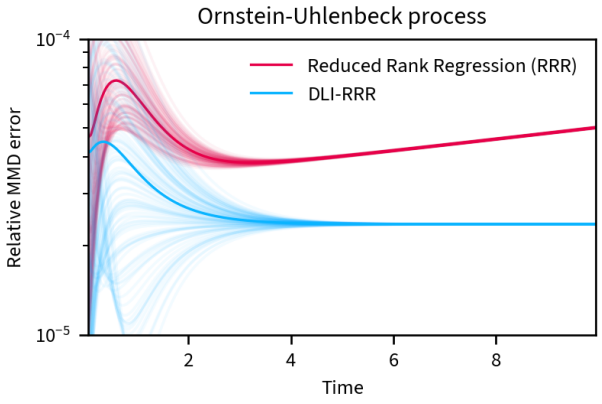

Figure 3. Distribution forecasting: Relative MMD error for the OU process for 100 independent experiments (thin lines).

Angles of Alanine Dipeptide We assess the forecasting performance of DLI estimators on a dataset of molecular dynamics simulations for the small molecule Alanine Dipeptide (Wehmeyer & Noé, 2018). The data comprises three independent 250 ns simulations, each containing records of the atomic positions, distances, and backbone dihedral angles. We train the estimators on 100 independent subsamples — each 5000 points long — from one of the provided trajectories. We have used the 45 pairwise atomic distances as input features and forecasted the two backbone dihedral angles, dubbed ϕ and ψ in the scientific community, over a forecast horizon of 0.1 ns. The angles ϕ and ψ are well known to encode the long-term behavior of the system, making them a perfect set of observables for the task of long-term forecasting. In Fig. 1 we report the forecasting Mean Absolute Error (MAE) for ϕ and ψ over a test set of 5000 points. The MAE has been computed using the minimum image convention, as angles are 2π-periodic observables. It is interesting to note how DLI estimators not only achieve a smaller error but also a significantly smaller variance across independent samplings of the training dataset. This is a key benefit of our methodology, as minimizing forecast uncertainty is crucial for strengthening risk management in strategic planning. We remark that, as usual, the forecasting error contains an *irreducible* component given by the intrinsic stochasticity of the process (see Appendix B.6). This additive component, however, is the same in both estimators.

8. Conclusions

In this paper, we have studied data-driven approaches for the long-term forecasting of ergodic discrete dynamical systems. These systems, which may be either deterministic or stochastic, are fully represented by the associated Koopman or transfer operator. We focused on the problem of predicting the conditional mean, conditional variance as well as the flow of state distributions from the initial one. Motivated by the observation that mainstream KOR estimators may fail at this task, we presented a conceptually simple and statistically principled approach which solves the above problem. Our theoretical analysis offers novel insights into the importance of estimator non-normality in long-term forecasting, contributing to a more comprehensive statistical learning theory for dynamical systems. A compelling feature of our method is its agnostic nature towards the KOR estimator. Moreover, it offers the advantages of low computational complexity and seamless integration with standard kernel scaling methods. In the future it would be interesting to extend our method and analysis to continuous-time systems, as well as tackle non-stationary processes.

Broader Impact

This paper presents work whose goal is to advance theoretical understanding of Machine learning techniques. There are several potential societal consequences of our work, none which we feel must be specifically highlighted here. I Our principled algorithms can help make machine learning

methods more reliable in practical scenarios.

References

Alexander, R. and Giannakis, D. Operator-theoretic framework for forecasting nonlinear time series with kernel analog techniques. *Physica D: Nonlinear Phenomena*, 409:132520, 2020.

Arbabi, H. and Mezić, I. Ergodic theory, dynamic mode decomposition, and computation of spectral properties of the koopman operator. *SIAM Journal on Applied Dynamical Systems*, 16(4):2096–2126, 2017.

Aronszajn, N. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

Azencot, O., Erichson, N. B., Lin, V., and Mahoney, M. Forecasting sequential data using consistent koopman autoencoders. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 475–485. PMLR, 2020.

Bandtlow, O. F. Estimates for norms of resolvents and an application to the perturbation of spectra. *Mathematische Nachrichten*, 267, 2004. URL https://api.semanticscholar.org/CorpusID:50535444.

Bevanda, P., Beier, M., Kerz, S., Lederer, A., Sosnowski, S., and Hirche, S. KoopmanizingFlows: Diffeomorphically Learning Stable Koopman Operators. *arXiv preprint arXiv.2112.04085*, 2021.

Bevanda, P., Beier, M., Lederer, A., Sosnowski, S., Hüllermeier, E., and Hirche, S. Koopman kernel regression. *arXiv preprint arXiv:2305.16215*, 2023.

Bouvrie, J. and Hamzi, B. Kernel Methods for the Approximation of Nonlinear Systems. *SIAM Journal on Control and Optimization*, 55(4):2460–2492, 2017. doi: 10.1137/14096815x. URL https://doi.org/10.1137/14096815x.

Briol, F.-X., Barp, A., Duncan, A. B., and Girolami, M. Statistical inference for generative models with maximum mean discrepancy. *arXiv preprint arXiv:1906.05944*, 2019.

Brunton, S. L., Budišić, M., Kaiser, E., and Kutz, J. N. Modern Koopman Theory for Dynamical Systems. *SIAM Review*, 64(2):229–340, 2022.

Budišić, M., Mohr, R., and Mezić, I. Applied Koopmanism. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 22(4):047510, 2012.

Caponnetto, A. and De Vito, E. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Cox, J. C., Ingersoll, J., and Ross, S. A. A theory of the term structure of interest rates. *Econometrica*, 53(2):385, 1985.

Da Prato, G. and Zabczyk, J. *Ergodicity for Infinite Dimensional Systems*. London Mathematical Society Lecture Note Series. Cambridge University Press, 1996.

Das, S. and Giannakis, D. Koopman spectra in reproducing kernel Hilbert spaces. *Applied and Computational Harmonic Analysis*, 49(2):573–607, 2020.

Davidchack, R. L., Ouldridge, T. E., and Tretyakov, M. V. New langevin and gradient thermostats for rigid body dynamics. *The Journal of Chemical Physics*, 142(14): 144114, 2015.

El-Fallah, O. and Ransford, T. Extremal growth of powers of operators satisfying resolvent conditions of kreiss-ritt type. *Journal of Functional Analysis*, 196(1):135–154, 2002. ISSN 0022-1236. doi: https://doi.org/10.1006/jfan.2002.3934. URL https://www.sciencedirect.com/science/article/pii/S0022123602939340.

Fan, F., Yi, B., Rye, D., Shi, G., and Manchester, I. R. Learning Stable Koopman Embeddings. *arXiv preprint arXiv.2110.06509*, 2021. doi: 10.48550/ARXIV.2110.06509. URL https://arxiv.org/abs/2110.06509.

Fischer, S. and Steinwart, I. Sobolev norm learning rates for regularized least-squares algorithms. *Journal of Machine Learning Research*, 21(205):1–38, 2020.

Hou, B., Sanjari, S., Dahlin, N., Bose, S., and Vaidya, U. Sparse learning of dynamical systems in RKHS: An operator-theoretic approach. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 13325–13352. PMLR, 2023.

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. doi: 10.1137/19M1305069. URL https://doi.org/10.1137/19M1305069.

Klus, S., Nüske, F., Koltai, P., Wu, H., Kevrekidis, I., Schütte, C., and Noé, F. Data-driven model reduction and transfer operator approximation. *Journal of Nonlinear Science*, 28(3):985–1010, 2018.

Klus, S., Schuster, I., and Muandet, K. Eigendecompositions of transfer operators in reproducing kernel Hilbert

spaces. *Journal of Nonlinear Science*, 30(1):283–315, 2019.

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. In *Advances in Neural Information Processing Systems*, 2022.

Kostic, V., Lounici, K., Novelli, P., and Pontil, M. Sharp spectral rates for koopman operator learning. In *Advances in Neural Information Processing Systems*, 2023a.

Kostic, V. R., Novelli, P., Grazzi, R., Lounici, K., and Pontil, M. Learning invariant representations of time-homogeneous stochastic dynamical systems. *arXiv:2307.09912*, 2023b.

Kutz, J. N., Brunton, S. L., Brunton, B. W., and Proctor, J. L. *Dynamic Mode Decomposition*. Society for Industrial and Applied Mathematics, 2016.

Lasota, A. and Mackey, M. C. *Chaos, Fractals, and Noise*, volume 97 of *Applied Mathematical Sciences*. Springer New York, 1994.

Li, Z., Meunier, D., Mollenhauer, M., and Gretton, A. Optimal rates for regularized conditional mean embedding learning. In *Advances in Neural Information Processing Systems*, 2022.

Lusch, B., Kutz, J. N., and Brunton, S. L. Deep learning for universal linear embeddings of nonlinear dynamics. *Nature Communications*, 9(1), 2018.

Meanti, G., Chatalic, A., Kostic, V. R., Novelli, P., Pontil, M., and Rosasco, L. Estimating koopman operators with sketching to provably learn large scale dynamical systems. In *Advances in Neural Information Processing Systems*, 2023.

Mezić, I. Analysis of fluid flows via spectral properties of the koopman operator. *Annual Review of Fluid Mechanics*, 45:357–378, 2013.

Mohan, N., Soman, K., and Kumar, S. S. A data-driven strategy for short-term electric load forecasting using dynamic mode decomposition model. *Applied Energy*, 232:229–244, 2018.

Morton, J., Witherden, F. D., Jameson, A., and Kochenderfer, M. J. Deep dynamical modeling and control of unsteady fluid flows. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 9278–9288, 2018.

Muandet, K., Fukumizu, K., Sriperumbudur, B., and Schölkopf, B. Kernel Mean Embedding of Distributions: A Review and Beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017. doi: 10.1561/2200000060. URL https://doi.org/10.1561/2200000060.

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.

Pascucci, A. *PDE and Martingale Methods in Option Pricing*. Springer Milan, 2011.

Proctor, J. L. and Eckhoff, P. A. Discovering dynamic patterns from infectious disease data using dynamic mode decomposition. *International health*, 7(2):139–145, 2015.

Saad, Y. *Numerical Methods for Large Eigenvalue Problems: Revised Edition*. Classics in Applied Mathematics, SIAM, 2011.

Scher, S. Toward data-driven weather and climate forecasting: Approximating a simple general circulation model with deep learning. *Geophysical Research Letters*, 45 (22):12–616, 2018.

Schütte, C., Huisinga, W., and Deuflhard, P. Transfer Operator Approach to Conformational Dynamics in Biomolecular Systems. In *Ergodic Theory, Analysis, and Efficient Simulation of Dynamical Systems*, pp. 191–223. Springer Berlin Heidelberg, 2001.

Steinwart, I. and Christmann, A. *Support Vector Machines*. Springer New York, 2008.

Trefethen, L. N. and Embree, M. *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*. Princeton University Press, 2020. ISBN 9780691213101. doi: doi:10.1515/9780691213101. URL https://doi.org/10.1515/9780691213101.

Wehmeyer, C. and Noé, F. Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics. *The Journal of Chemical Physics*, 148 (24):241703, June 2018. doi: 10.1063/1.5011399. URL https://doi.org/10.1063/1.5011399.

Williams, M. O., , Rowley, C. W., and Kevrekidis, I. G. A kernel-based method for data-driven Koopman spectral analysis. *Journal of Computational Dynamics*, 2(2):247–265, 2015.

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

Zabczyk, J. *Mathematical Control Theory: An Introduction*. Systems & Control: Foundations & Applications. Springer International Publishing, 2020. ISBN

9783030447762. URL https://books.google.it/books?id=45tfzQEACAAJ.

Supplementary Material

The appendix is organized as follows:

- Appendix A provides a summary of important notations in Table 2, additional background on the Koopman operator framework and the detailed algorithm (Alg. 1) used to implement the DLI paradigm.
- Appendix B contains the complete proofs of the results in the main body of the paper. Specifically,
  - Appendices B.1-B.4 cover the extension of the existing bounds to the case of centered features,
  - Appendix B.5 proves the result for RRR estimator,
  - Appendix B.6 proves the bounds for the conditional variance forecasting
  - Appendix B.7 presents the extension to non-iid sampling via β-mixing.
- Appendix C contains additional details on the experiments we performed in the main body of the paper.

A. Background

A.1. Markov Transfer Operators

Let $\mathbf{X} := \{X_t \colon t \in \mathbb{N}\}$ be a family of random variables with values in a measurable space $(\mathcal{X}, \Sigma_{\mathcal{X}})$, called state space. We call $\mathbf{X}$ a *Markov chain* if $\mathbb{P}\{X_{t+1} \in B \,|\, X_{[t]}\} = \mathbb{P}\{X_{t+1} \in B \,|\, X_t\}$. Further, we call $\mathbf{X}$ *time-homogeneous* if there exists $p \colon \mathcal{X} \times \Sigma_{\mathcal{X}} \to [0,1]$, called *transition kernel*, such that, for every $(x, B) \in \mathcal{X} \times \Sigma_{\mathcal{X}}$ and every $t \in \mathbb{N}$,

$$\mathbb{P}\left\{X_{t+1} \in B | X_t = x\right\} = p(x, B).$$

A large class of Markov chains consists of those endowed with *invariant measure*, denoted as π, that satisfies the equation $\pi(B) = \int_{\mathcal{X}} \pi(dx) p(x, B)$, $B \in \Sigma_{\mathcal{X}}$, see e.g. (Da Prato & Zabczyk, 1996). For such cases, we can consider the space of square-integrable functions on $\mathcal{X}$ relative to the measure π, denoted as $L^2_\pi(\mathcal{X})$, and define the *Markov transfer operator*, $A_\pi \colon L^2_\pi(\mathcal{X}) \to L^2_\pi(\mathcal{X})$

$$[A_\pi f](x) := \int_{\mathcal{X}} p(x, dy) f(y) = \mathbb{E}\left[f(X_{t+1}) | X_t = x\right], \quad f \in L^2_\pi(\mathcal{X}),\, x \in \mathcal{X}. \tag{32}$$

Since it easy to see that $\|A_\pi f\| \leq \|f\|$, we conclude that $\|A_\pi\| \leq 1$, i.e. the Markov transfer operator is a bounded linear operator. Moreover, recalling that $\mathbb{1}_\pi(x) = 1$ for π-a.e. $x \in \mathcal{X}$, since $A_\pi \mathbb{1}_\pi = \mathbb{1}_\pi$, we see that $1 \leq \rho(A_\pi) \leq \|A_\pi\| \leq 1$, i.e. $\rho(A_\pi) = \|A_\pi\| = 1$.

A.2. Koopman Mode Decomposition (KMD)

In dynamical systems, A_π is known as the (stochastic) *Koopman operator* on the space of observables $\mathcal{F} = L^2_\pi(\mathcal{X})$. An essential characteristic of this operator is its linearity, which can be harnessed for the computation of a spectral decomposition. Indeed, in many situations, especially when dealing with compact Koopman operators, there exist complex scalars $\lambda_i \in \mathbb{C}$ and observables $\psi_i \in L^2_\pi(\mathcal{X})$ that satisfy the eigenvalue equation $A_\pi \psi_i = \lambda_i \psi_i$. Leveraging the eigenvalue decomposition, the dynamical system can be decomposed into superposition of simpler signals that can be used in different tasks such as system identification and control, see e.g. (Brunton et al., 2022). More precisely, given an observable $f \in \operatorname{span}\{\psi_i \,|\, i \in \mathbb{N}\}$ there exist corresponding scalars $\gamma_i^f \in \mathbb{C}$ known as Koopman modes of f, such that

$$A_\pi^t f(x) = \mathbb{E}[f(X_t) \,|\, X_0 = x] = \sum_{i \in \mathbb{N}} \lambda_i^t \gamma_i^f \psi_i(x), \quad x \in \mathcal{X},\, t \in \mathbb{N}. \tag{33}$$

This formula is known as *Koopman Mode Decomposition* (KMD) (Budišić et al., 2012; Arbabi & Mezić, 2017). It decomposes the expected dynamics observed by f into *stationary* modes γ_i^f that are combined with *temporal changes* governed by eigenvalues λ_i and *spatial changes* governed by the eigenfunctions ψ_i. We notice however that the Koopman operator, in general, is not a normal compact operator, hence its eigenfunctions may not form a complete orthonormal basis of the space which makes learning KMD challenging.

notation	meaning
μ_t	law of the state of process at the time t
q_t	density of the law of the state of process at the time t w.r.t. invariant distribution
$k(\cdot,\cdot)$	symmetric positive definite kernel function
$\phi(x)$	canonical feature map associated to $x \in \mathcal{X}$ also denoted by k_x
k_ν	kernel mean embedding of the measure ν
A_π	Koopman operator
S_π	canonical injection of $\mathcal{H} \hookrightarrow L^2_\pi(\mathcal{X})$
$A_\pi S_\pi$	restriction of the Koopman operator to $\mathcal{H}$
$\mathbb{1}_\pi$	function in $L^2_\pi(\mathcal{X})$ with the constant output 1
J_π	projection onto $\mathbb{1}^\perp$ in $L^2_\pi(\mathcal{X})$
$\mathbf{A}_\pi$	deflated Koopman operator
$\mathbf{S}_\pi$	projected canonical injection of $\mathcal{H} \hookrightarrow L^2_\pi(\mathcal{X})$
$\mathbf{A}_\pi\mathbf{S}_\pi$	restriction of the deflated Koopman operator to $\mathcal{H}$
$\mathcal{R}$	true risk
$\mathcal{R}^\circ$	true centered risk
$\widehat{\mathcal{R}}$	empirical risk
$\widehat{\mathcal{R}}^\circ$	empirical centered risk
$\mathcal{E}$	true error
$\widehat{\mu}_t$	empirical (signed) measure that estimates μ_t at time $t \in \mathbb{N}$
$\widehat{h}_t$	function in $\mathcal{H}$ that estimates $\mathbb{E}[h(X_t) \mid X_0 = \cdot]$
$\mathcal{E}^\circ$	true centered error
$\mathbb{1}_n$	normalized constant vector in $\mathbb{R}^n$ with all components equal to $1/\sqrt{n}$
J_n	projection onto $\text{span}(\mathbb{1}_n)^\perp$ in $\mathbb{R}^n$
$\widehat{S}_x$	sampling operator of the inputs
$\widehat{S}_y$	sampling operator of the outputs
C	covariance operator
$\widehat{C}$	empirical covariance operator
T	cross-covariance operator
$\mathbf{C}$	centered covariance operator
$\widehat{\mathbf{C}}$	centered empirical covariance operator
$\widehat{T}$	empirical cross-covariance operator
$\mathbf{T}$	centered cross-covariance operator
$\widehat{\mathbf{T}}$	centered empirical cross-covariance operator
K_x	input kernel matrix
K_y	output kernel Gramm matrix
K_γ	regularized input kernel matrix
$\mathbf{K}_x$	centered input kernel matrix
$\mathbf{K}_y$	centered output kernel Gramm matrix
$\mathbf{K}_\gamma$	regularized centered input kernel matrix
$\mathcal{B}(\mathcal{H})$	the set of bounded operators on $\mathcal{H}$
$\mathcal{B}_r(\mathcal{H})$	the set of operators on $\mathcal{H}$ of a finite rank at most r
$\text{HS}(\mathcal{H})$	the set of Hilbert-Schmidt (HS) operators on $\mathcal{H}$
$\text{Sp}(\cdot)$	spectrum of a bounded operator
$\rho(\cdot)$	spectral radius of a bounded operator
$d(\cdot)$	distance to instability of a bounded operator
$\eta(\cdot)$	Kreiss constant of a bounded operator

Table 2. Summary of used notations.

A.3. Kernel Based Learning of Koopman Operators

In many practical situations, the Koopman operator A_π is unknown, but data from one or multiple system trajectories are available. A learning framework called Koopman Operator Regression (KOR) was introduced in (Kostic et al., 2022) to estimate the Koopman operator A_π on $L^2_\pi(\mathcal{X})$ using reproducing kernel Hilbert spaces (RKHS). More precisely, consider an RKHS denoted as $\mathcal{H}$ with kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (Aronszajn, 1950). Let $\phi : \mathcal{X} \to \mathcal{H}$ be an associated feature map such that $k(x, y) = \langle \phi(x), \phi(y) \rangle$ for all $x, y \in \mathcal{X}$. We assume that $k(x, x) \leq c_{\mathcal{H}} < \infty$, π- almost surely. This ensures that $\mathcal{H} \subseteq L^2_\pi(\mathcal{X})$, and the injection operator $S_\pi \colon \mathcal{H} \to L^2_\pi(\mathcal{X})$, defined as $(S_\pi f)(x) = f(x)$ for $x \in \mathcal{X}$, along with its adjoint $S^*_\pi \colon L^2_\pi(\mathcal{X}) \to \mathcal{H}$ are well-defined Hilbert-Schmidt operators (Caponnetto & De Vito, 2007; Steinwart & Christmann, 2008). Then, the Koopman operator, when restricted to $\mathcal{H}$, is given by

$$A_\pi S_\pi \colon \mathcal{H} \to L^2_\pi(\mathcal{X}).$$

Unlike A_π, the operator $A_\pi S_\pi$ is Hilbert-Schmidt, which allows us to estimate $A_\pi S_\pi$ by minimizing the following *risk*

$$\mathcal{R}(G) = \mathbb{E}_{x\sim\pi} \sum_{i\in\mathbb{N}} \mathbb{E}\big[(h_i(X_{t+1}) - (Gh_i)(X_t))^2 \,|X_t = x\big] \tag{34}$$

over Hilbert-Schmidt operators $G \in \mathrm{HS}(\mathcal{H})$, where $(h_i)_{i\in\mathbb{N}}$ is an orthonormal basis of $\mathcal{H}$. Moreover, we can write down a bias-variance decomposition of the risk $\mathcal{R}(G) = \mathcal{R}_0 + \mathcal{E}_{\mathrm{HS}}(G)$, where

$$\mathcal{R}_0 = \|S_\pi\|^2_{\mathrm{HS}} - \|A_\pi S_\pi\|^2_{\mathrm{HS}} \geq 0 \;\text{ and }\; \mathcal{E}_{\mathrm{HS}}(G) = \|A_\pi S_\pi - S_\pi G\|^2_{\mathrm{HS}}, \tag{35}$$

are the irreducible risk (i.e. the variance term in the classical bias-variance decomposition) and the excess risk, respectively. This can be equivalently expressed in the terms of embedded dynamics in RKHS as:

$$\underbrace{\mathbb{E}_{(X,Y)}\|\phi(Y) - G^*\phi(X)\|^2}_{\mathcal{R}(G)} = \underbrace{\mathbb{E}_{(X,Y)}\|g_p(X) - \phi(Y)\|^2}_{\mathcal{R}_0} + \underbrace{\mathbb{E}_{X\sim\pi}\|g_p(X) - G^*\phi(X)\|^2}_{\mathcal{E}_{\mathrm{HS}}(G)}, \tag{36}$$

where (X, Y) is has the joint probability measure of two consecutive states of the Markov chain, and the regression function $g_p \colon \mathcal{X} \to \mathcal{H}$ is defined as $g_p(x) := \mathbb{E}[\phi(X_{t+1}) \,|\, X_t = x] = \int_{\mathcal{X}} p(x, dy)\phi(y)$, $x \in \mathcal{X}$, and is known as the *conditional mean embedding* (CME) of the conditional probability p into $\mathcal{H}$. It was also shown that using universal kernels one can approximate the restriction of Koopman arbitrary well, i.e. excess risk can be made arbitrarily small $\inf_{G\in\mathrm{HS}(\mathcal{H})} \mathcal{E}_{\mathrm{HS}}(G) = 0$.

Therefore, to develop estimators one can consider the problem of minimizing the Tikhonov regularized risk

$$\min_{G\in\mathrm{HS}(\mathcal{H})} \mathcal{R}^\gamma(G) := \mathcal{R}(G) + \gamma\|G\|^2_{\mathrm{HS}}, \tag{37}$$

where $\gamma > 0$. Denoting the covariance matrix as $C := S^*_\pi S_\pi = \mathbb{E}_{X\sim\pi}\phi(X) \otimes \phi(X)$ and cross-covariance matrix $T := S^*_\pi A_\pi S_\pi = \mathbb{E}_{(X,Y)}\phi(X) \otimes \phi(Y)$, and regularized covariance as $C_\gamma := C + \gamma I_{\mathcal{H}}$, one easily shows that $G_\gamma := C^{-1}_\gamma T$ is the unique solution of (37) which is known as the Kernel Ridge Regression (KRR) estimator of A_π.

Low rank estimators of the Koopman operator have also been considered. Notably, Principal Component Regression (PCR) estimator given by $[\![C]\!]^\dagger_r T$, where $[\![\cdot]\!]_r$ denotes the r-truncated SVD of the Hilbert-Schmidt operator. However, it is observed that both KRR and PCR estimators can fail in estimating well the leading Koopman eigenvalues (Kostic et al., 2023a). To mitigate this, Reduced Rank Regression (RRR) estimator has been introduced in (Kostic et al., 2022) as the optimal one that solves (37) with an additional rank constraint by minimizing over the class of rank-r HS operators $\mathcal{B}_r(\mathcal{H}) := \{G \in \mathrm{HS}(\mathcal{H}) \mid \mathrm{rank}(G) \leq r\}$, where $1 \leq r < \infty$, i.e.

$$C^{-1/2}_\gamma [\![C^{-1/2}_\gamma T]\!]_r = \operatorname*{arg\,min}_{G\in\mathcal{B}_r(\mathcal{H})} \mathcal{R}^\gamma(G). \tag{38}$$

Now, assuming that data $\mathcal{D} = \{(x_i, y_i)\}_{i\in[n]}$ is collected, the estimators are typically obtained via the regularized empirical risk $\widehat{\mathcal{R}}^\gamma(G) := \frac{1}{n}\sum_{i\in[n]}\|\phi(y_i) - G^*\phi(x_i)\|^2 + \gamma\|G\|^2_{\mathrm{HS}}$ minimization (RERM). Introducing the sampling operators for data $\mathcal{D}$ and RKHS $\mathcal{H}$ by

$$\widehat{S}_x \colon \mathcal{H} \to \mathbb{R}^n \quad \text{s.t. } f \mapsto \tfrac{1}{\sqrt{n}}[f(x_i)]_{i\in[n]} \quad \text{and} \qquad \widehat{S}_y \colon \mathcal{H} \to \mathbb{R}^n \quad \text{s.t. } f \mapsto \tfrac{1}{\sqrt{n}}[f(y_i)]_{i\in[n]},$$

and their adjoints by

$$\widehat{S}_x^*\colon \mathbb{R}^n \to \mathcal{H} \quad \text{s.t. } w \mapsto \tfrac{1}{\sqrt{n}} \sum_{i\in[n]} w_i \phi(x_i) \quad \text{and} \qquad \widehat{S}_y^*\colon \mathbb{R}^n \to \mathcal{H} \quad \text{s.t. } w \mapsto \tfrac{1}{\sqrt{n}} \sum_{i\in[n]} w_i \psi(y_i),$$

we obtain $\widehat{\mathcal{R}}^\gamma(G) = \|\widehat{S}_y - \widehat{S}_x G\|_{\text{HS}}^2 + \gamma \|G\|_{\text{HS}}^2$.

In the following we also use the empirical covariance operator defined as

$$\widehat{C} := \widehat{S}_x^* \widehat{S}_x = \tfrac{1}{n} \sum_{i\in[n]} \phi(x_i) \otimes \phi(x_i), \tag{39}$$

and the empirical cross-covariance operator

$$\widehat{T} := \widehat{S}_x^* \widehat{S}_y = \tfrac{1}{n} \sum_{i\in[n]} \phi(x_i) \otimes \phi(y_i). \tag{40}$$

Additionally, we let $\widehat{C}_\gamma := \widehat{C} + \gamma I_{\mathcal{H}}$ be the regularized empirical covariance. Then we obtain the empirical estimators of the Koopman operator on an RKHS that correspond to the population ones: empirical KRR estimator $\widehat{G}_\gamma := \widehat{C}_\gamma^{-1}\widehat{T}$, empirical PCR estimator $[\![\widehat{C}]\!]_r^\dagger \widehat{T}$, and empirical RRR estimator $\widehat{C}_\gamma^{-1/2} [\![\widehat{C}_\gamma^{-1/2}\widehat{T}]\!]_r$.

Noting that all of the empirical estimators above are of the form $\widehat{G} = \widehat{S}_x U_r V_r^\top \widehat{S}_y$, where $U_r, V_r \in \mathbb{R}^{n\times r}$ and $r \in [n]$ which are computed using (normalized) kernel Gramm matrices $K_x := \widehat{S}_x \widehat{S}_x^* = \frac{1}{n}[k(x_i, x_j)]_{i,j\in[n]}$ and $K_y := \widehat{S}_y \widehat{S}_y^* = \frac{1}{n}[k(y_i, y_j)]_{i,j\in[n]}$, see (Kostic et al., 2022). In the next section (especially in Theorem A.2 and Alg. 1) we explain how to compute these estimators in practice within the proposed DLI framework.

A.4. Deflate-Learn-Inflate Estimation of Koopman Operator

1) Deflation: Projecting the Koopman Operator

This step consists in projecting the Koopman operator onto the subspace of $L^2_\pi(\mathcal{X})$ orthogonal to $\mathbb{1}_\pi$: the operator to learn is no longer A_π but $A_\pi - \mathbb{1}_\pi \otimes \mathbb{1}_\pi$. Denote by $J_\pi = I - \mathbb{1}_\pi \otimes \mathbb{1}_\pi$ the orthogonal projector onto $\text{span}(\mathbb{1}_\pi)^\top$ in $L^2_\pi(\mathcal{X})$, and two operators J_π and A_π commute since $\mathbb{1}$ is a left and right singular function of A_π. Hence, we have that $A_\pi - \mathbb{1}_\pi \otimes \mathbb{1}_\pi = A_\pi J_\pi = J_\pi A_\pi = J_\pi A_\pi J_\pi$, which we denote by $\mathbf{A}_\pi$, which implies that the restriction of the deflated Koopman operator to $\mathcal{H}$, i.e. $\mathbf{A}_\pi S_\pi$ can be written as $\mathbf{A}_\pi \mathbf{S}_\pi = J_\pi A_\pi S_\pi$, where $\mathbf{S}_\pi := J_\pi S_\pi$ is projected injection operator.

Remark A.1. In the specific context of Koopman regression the process of deflation is equivalent to centering the feature map. Indeed deflation leads to learning the Koopman operator projected onto the $L^2_\pi(\mathcal{X})$-orthogonal subspace of the constant function, that is zero mean functions. On the sample level, this means substracting the empirical mean of the data set, a procedure which is therefore equivalent to centering the feature map. Consequently, the following procedure is motivated by the theory of centering feature maps.

2) Learn: Compute Estimators for $A_\pi J_\pi$

Learning $A_\pi J_\pi$ is the most interesting part of the Deflate-Learn-Inflate process, based on regression techniques. Previous literature on Koopman learning Kostic et al. (2022; 2023a); Li et al. (2022); Klus et al. (2019; 2018) provides three popular estimators based on a statistics approach which we proceed to introduce. Since the regression problem is now learning the projected Koopman operator $A_\pi J_\pi$ we will hereby, analogously to (35), introduce the risk associated to this learning problem for a Hilbert-Schmidt operator $\mathbf{G} \in \text{HS}(\mathcal{H})$ as $\mathcal{R}^\circ(G) = \mathcal{R}_0^\circ + \mathcal{E}_{\text{HS}}^\circ(\mathbf{G})$, where

$$\mathcal{R}_0^\circ := \|J_\pi S_\pi\|_{\text{HS}}^2 - \|J_\pi A_\pi S_\pi\|_{\text{HS}}^2 = \|\mathbf{S}_\pi\|_{\text{HS}}^2 - \|\mathbf{A}_\pi \mathbf{S}_\pi\|_{\text{HS}}^2 \geq 0, \tag{41}$$

is the centered irreducible risk (i.e. the variance term in the classical bias-variance decomposition) and

$$\mathcal{E}_{\text{HS}}^\circ(G) := \|J_\pi A_\pi S_\pi - J_\pi S_\pi \mathbf{G}\|_{\text{HS}}^2 = \|\mathbf{A}_\pi \mathbf{S}_\pi - \mathbf{S}_\pi \mathbf{G}\|_{\text{HS}}^2, \tag{42}$$

is the centered excess risk. After some algebra, similarly as in (Kostic et al., 2022), centered risk can be equivalently expressed as:

$$\mathcal{R}^\circ(\mathbf{G}) = \mathbb{E}_{(X,Y)} \|(k_Y - k_\pi) - \mathbf{G}^*(k_X - k_\pi)\|^2. \tag{43}$$

The following section explains how an estimator $\mathbf{S}_\pi\mathbf{G}$ for the Hilbert-Schmidt operator $\mathbf{A}_\pi\mathbf{S}_\pi$ is computed using regression algorithms. Given a dataset $\mathcal{D} := (x_i, y_i)_{i=1}^n$ of sampled consecutive states, we introduce the empirical mean square error of an estimator $\mathbf{G}$ for $\mathbf{A}_\pi\mathbf{S}_\pi$:

$$\widehat{\mathcal{R}}^\circ(\mathbf{G}) = \frac{1}{n}\sum_{i=1}^n \left\| \left(\phi(y_i) - \frac{1}{n}\sum_{i=1}^n \phi(y_i)\right) - \mathbf{G}^* \left(\phi(x_i) - \frac{1}{n}\sum_{i=1}^n \phi(x_i)\right)\right\|_{\mathrm{HS}}^2 \tag{44}$$

which is merely the empirical version of 43. the excess risk associated to an estimator $\mathbf{G}$ for $\mathbf{A}_\pi\mathbf{S}_\pi$. Notice that here we are centering the feature map on the input and output space (by removing the mean). We also introduce the regularised risk which defines two popular estimators:

$$\widehat{\mathcal{R}}^\circ_\gamma(\mathbf{G}) = \widehat{\mathcal{R}}^\circ(\mathbf{G}) + \gamma\|\mathbf{G}\|^2_{HS}. \tag{45}$$

The framework of statistical learning translates the approximation problem to a minimisation problem on the space of Hilbert-Schmidt operator acting on $\mathcal{H}$. This formulation was not introduced in the setting of Koopman operator regression until (Kostic et al., 2022) and was key to giving some statistical insight onto the different learning techniques. The paper gives various statistical properties of three supervised learning algorithms: Kernel Ridge Regression (KRR), Principal Component Regression (PCR) and Reduced Rank Regression (RRR). The KRR estimator minimises the empirical regularised risk (45) whilst the RRR algorithm minimises the same regularised risk under a fixed rank constraint. On the other hand the PCR estimator does not minimise the empirical risk but projects the output on the leading r eigenvectors of the covariance operator, that is the vectors responsible for the most variability of the inputs.

All of these estimators can be expressed in a unified form. Namely, denoting the normalized constant vector in $\mathbb{R}^n$ by $\mathbb{1}_n := n^{-1/2}[1, \dots, 1]^T$ and the orthogonal projector to orthogonal complement by $J_n := I_n - \mathbb{1}_n \otimes \mathbb{1}_n$, we can introduce the projected sampling operators $\widehat{\mathbf{S}}_x := J_n\widehat{S}_x$ and $\widehat{\mathbf{S}}_y := J_n\widehat{S}_y$, which, recalling that $k_\pi = \mathbb{E}_{X\sim\pi}\phi(X)$, $\hat{\pi}_x = \frac{1}{n}\sum_{i\in[n]}\phi(x_i)$ and $\hat{\pi}_y = \frac{1}{n}\sum_{i\in[n]}\phi(y_i)$, are used to define empirical versions

$$\widehat{\mathbf{C}} := \widehat{\mathbf{S}}_x^*\widehat{\mathbf{S}}_x = \sum_{i\in[n]} (\phi(x_i) - k_{\hat{\pi}_x}) \otimes (\phi(x_i) - k_{\hat{\pi}_x}) \text{ and } \widehat{\mathbf{T}} := \widehat{\mathbf{S}}_x^*\widehat{\mathbf{S}}_y = \sum_{i\in[n]} (\phi(x_i) - k_{\hat{\pi}_x}) \otimes \left(\phi(y_i) - k_{\hat{\pi}_y}\right),$$

of the centered covariance matrix as $\mathbf{C} := \mathbf{S}_\pi^*\mathbf{S}_\pi = \mathbb{E}_{X\sim\pi}(\phi(X) - k_\pi) \otimes (\phi(X) - k_\pi)$ and centered cross-covariance matrix $\mathbf{T} := \mathbf{S}_\pi^*\mathbf{A}_\pi\mathbf{S}_\pi = \mathbb{E}_{(X,Y)}(\phi(X) - k_\pi) \otimes (\phi(Y) - k_\pi)$, respectively. Furthermore, we can introduce centered Kernel matrices associated to the input points:

$$\mathbf{K}_x := J_nK_xJ_n = K_x - (K_x\mathbb{1}_n)\mathbb{1}_n^\top - \mathbb{1}_n(K_x\mathbb{1}_n)^\top + (\mathbb{1}_n^\top K_x\mathbb{1}_n)\mathbb{1}_n\mathbb{1}_n^\top \tag{46}$$

and, analogously, the one associated to the output points: $\mathbf{K}_y := J_nK_yJ_n$. Then, a unified form for centered empirical estimators is: $\mathbf{G} = \widehat{\mathbf{S}}_x^*W\widehat{\mathbf{S}}_y = \widehat{S}_x^*J_nWJ_n\widehat{S}_y$, where W is a square matrix of size n (the number of samples), which we will refer to as the matrix form of the estimator from now on. The following theorem gives the expression of the matrix form of the estimators derived by the previously discussed algorithms. The proofs closely follow those presented in (Kostic et al., 2022) for the estimators of the Koopman operator A_π. The reader is encouraged to read this paper which introduces the empirical risk minimisation problem.

Theorem A.2. *The deflated Koopman operator restricted to the RKHS $\mathcal{H}$, i.e. $\mathbf{A}_\pi\mathbf{S}_\pi = \mathbf{A}_\pi S_\pi$, can be empirically estimated by $\widehat{\mathbf{G}} = \widehat{\mathbf{S}}_x^*W\widehat{\mathbf{S}}_y$, where $W \in \mathbb{R}^{n\times n}$ is determined as follows:*

(i) The Kernel Ridge Regression (KRR) algorithm yields $\widehat{\mathbf{G}} = \widehat{\mathbf{G}}_\gamma := \widehat{\mathbf{C}}_\gamma^{-1}\widehat{\mathbf{T}}$ and $W = (\mathbf{K}_x + \gamma I_n)^{-1}$.

(ii) The Principal Component Regression algorithm (PCR) yields $\widehat{\mathbf{G}} = \widehat{\mathbf{G}}^{\mathrm{PCR}}_{r,\gamma} := [\![\widehat{\mathbf{C}}]\!]_r^\dagger\widehat{\mathbf{T}}$ and $W = U_rV_r^\top$, where $[\![\mathbf{K}_x]\!]_r = V_r\Sigma_rV_r^\top$ is the r-trunacted SVD of $\mathbf{K}_x$ and $U_r := V_r\Sigma_r^\dagger$.

(iii) The Reduced Rank Regression algorithm (RRR) yields $\widehat{\mathbf{G}} = \widehat{\mathbf{G}}^{\mathrm{RRR}}_{r,\gamma} := \widehat{C}_\gamma^{-1/2}[\![\widehat{C}_\gamma^{-1/2}\widehat{T}]\!]_r$ and $W = U_rV_r^\top$, where $V_r := \mathbf{K}_xU_r$ and $U_r = [u_1|\dots|u_r] \in \mathbb{R}^{n\times r}$ is such that (σ_i, u_i) are the solutions to the generalised eigenvalue problem :

$$\mathbf{K}_y\mathbf{K}_xu_i = \sigma_i^2(\mathbf{K}_x + \gamma I_n)u_i \text{ normalised such that } u_i^\top\mathbf{K}_x(\mathbf{K}_x + \gamma I_n)u_i = 1.$$

3) Inflation: Preservation of Probability Mass

Finally, we use $\widehat{\mathbf{G}}$ to obtain the empirical estimates of $\mathbb{E}[h(X_t) \,|\, X_0 = \cdot]$, $h \in \mathcal{H}$, and μ_t, for all $t \in \mathbb{N}$, by putting back the leading eigenpair that we have removed during the deflate step. Recalling (21) and (31), respectively, this results in

$$\widehat{h}_t(x) := [\widehat{\mathbf{G}}^t h](x) + \langle k_{\hat{\pi}_y}, h\rangle - \langle k_{\hat{\pi}_x}, \widehat{\mathbf{G}}^t h\rangle, x \in \mathcal{X}, \text{ and } k_{\widehat{\mu}_t} = k_{\hat{\pi}_y} + (\widehat{\mathbf{G}}^*)^t (k_{\widehat{\mu}_0} - k_{\hat{\pi}_y})$$

where $\widehat{\mu}_0 = n_0^{-1} \sum_{i\in[n_0]} \delta_{z_i}$ is the initial empirical measure. Thus, recalling (22), (23), and using Theorem A.2 and $w_t^0 = \frac{1}{\sqrt{n}} W_t K_{xz} \mathbb{1}_{n_0}$ leads to Algorithm 1 that results in the sequence empirical measures $\widehat{\mu}_t = \sum_{i\in[n]} m_{t,i} \delta_{y_i}$ supported on the output points $(y_i)_{i\in[n]}$, which by construction satisfy $\widehat{\mu}_t(\mathcal{X}) = 1$ since $\sum_{j\in[n]} m_{t,j} = \mathbb{1}_n^\top \mathbb{1}_n = 1$ due to $J_n \mathbb{1}_n = 0$.

Concerning the observables, as shown in (29)-(30) these two estimators are related via

$$\langle k_{\widehat{\mu}_0}, \widehat{h}_t\rangle = \langle k_{\widehat{\mu}_t}, h\rangle,$$

and, hence we can forecast the observable as

$$\widehat{h}_t(x) = \langle k_z, \widehat{h}_t\rangle = \sum_{i\in[n]} m_{t,i} h(y_i)$$

simply setting $n_0 = 1$ and $z_1 = x$ in the algorithm bellow.

Algorithm 1 Forecasting observables and measures with KRR/PCR/RRR estimator via DLI framework

Require: Dataset $\mathcal{D}_n = (x_i, y_i)_{i\in[n]}$ from the process in stationary regime and samples $(z_i)_{i\in[n_0]}$ from some initial measure μ_0; hyperparameters $\gamma > 0$ and/or $r \in [n]$; forecasting horizon $T \in \mathbb{N}$.
- **if** $r = n$ **then** {*KRR estimator*}
  - Solve $(\mathbf{K}_x + \gamma I)\widetilde{m}_0 = J_n(K_{xz}\mathbb{1}_{n_0} - K_{xy}\mathbb{1}_n)$ in $\widetilde{m}_0$
  - Update $\widetilde{m}_0 \leftarrow J_n \widetilde{m}_0$
  - **for** $t = 1, \ldots, T-1$ **do**
    - Compute $m_t \leftarrow (\mathbb{1}_n + \widetilde{m}_{t-1})/\sqrt{n}$
    - Solve $(\mathbf{K}_x + \gamma I)\widetilde{w}_t = J_n K_{xy} \widetilde{m}_{t-1}$ in $\widetilde{m}_t$
    - Update $\widetilde{m}_t \leftarrow J_n \widetilde{m}_t$
  - **end for**
  - Compute $m_T \leftarrow (\mathbb{1}_n + \widetilde{m}_{T-1})/\sqrt{n}$
- **else** {*low rank estimators*}
  - **if** $\gamma = 0$ **then** {*PCR estimator*}
    - Compute $U_r, V_r \in \mathbb{R}^{n\times r}$ using Theorem A.2(ii)
  - **else** {*RRR estimator*}
    - Compute $U_r, V_r \in \mathbb{R}^{n\times r}$ using Theorem A.2(iii)
  - **end if**
  - Update $U_r \leftarrow J_n U_r$ and $V_r \leftarrow J_n V_r$
  - Compute $\widetilde{m}_0 \leftarrow U_r^\top (K_{xz}\mathbb{1}_{n_0} - K_{xy}\mathbb{1}_n)$
  - Compute $M \leftarrow U_r^\top K_{xy} V_r$
  - **for** $t = 1, \ldots, T-1$ **do**
    - Compute $m_t \leftarrow (\mathbb{1}_n + V_r^\top \widetilde{m}_{t-1})/\sqrt{n}$
    - Update $\widetilde{m}_t \leftarrow M \widetilde{m}_{t-1}$
  - **end for**
  - Compute $m_T \leftarrow (\mathbb{1}_n + V_r^\top \widetilde{m}_{T-1})/\sqrt{n}$
- **end if**

Ensure: Vectors of weights $m_t \in \mathbb{R}^n$ that define empirical measures $\widehat{\mu}_t = \sum_{i\in[n]} m_{T,i} \delta_{y_i}$, $t \in [T]$.

B. Proofs of Main Results

B.1. Main Assumptions

The following assumptions were used in Sec. 5 to derive the learning bounds:

(BK) *Boundedness.* There exists $c_{\mathcal{H}} > 0$ such that $\operatorname*{ess\,sup}_{x\sim\pi} k(x,x) \leq c_{\mathcal{H}}$, i.e. $\phi \in L^\infty_\pi(\mathcal{X},\mathcal{H})$.

(RC) *Regularity condition.* For some $\alpha \in [1,2]$ there exists $a > 0$ such that $TT^* \preceq a^2 C^{1+\alpha}$, with $T = S^*_\pi A_\pi S_\pi$.

(RC*) *Regularity condition on the deflated operator* For some $\alpha \in [1,2]$ there exists $a > 0$ such that $\mathbf{T}\mathbf{T}^* \preceq a^2 \mathbf{C}^{1+\alpha}$.

(SD) *Spectral Decay.* There exists $\beta \in (0,1]$ and a constant $b > 0$ such that $\lambda_j(C) \leq b\, j^{-1/\beta}$, for all $j \in J$.

(SD*) *Spectral Decay of the centered operator.* There exists $\beta \in (0,1]$ and a constant $b > 0$ such that $\lambda_j(\mathbf{C}) \leq b\, j^{-1/\beta}$, for all $j \in J$.

We start by observing that $S_\pi \in \mathrm{HS}\left(\mathcal{H}, L^2_\pi(\mathcal{X})\right)$, and, hence $\mathbf{S}_\pi \in \mathrm{HS}\left(\mathcal{H}, L^2_\pi(\mathcal{X})\right)$, too. Hence, according to the spectral theorem for positive self-adjoint operators, has an SVD, i.e. there exists at most countable positive sequence $(\sigma_j)_{j\in N}$, where $N := \{1,2,\dots,\} \subseteq \mathbb{N}$, and ortho-normal systems $(\ell_j)_{j\in N}$ and $(h_j)_{j\in N}$ of $\mathrm{cl}(\mathrm{Im}(\mathbf{S}_\pi))$ and $\mathrm{Ker}(\mathbf{S}_\pi)^\perp$, respectively, such that $\mathbf{S}_\pi h_j = \sigma_j \ell_j$ and $\mathbf{S}^*_\pi \ell_j = \sigma_j h_j$, $j \in N$. Moreover, since $\mathrm{cl}(\mathrm{Im}(\mathbf{S}_\pi)) \subseteq \mathrm{Im}(J_\pi)$, we also have $J_\pi \ell_j = \ell_j$, i.e. $\mathbb{E}_{X\sim\pi}[\ell_j(X)] = 0$, $j \in N$.

Now, given $\alpha \geq 0$, let us define scaled injection operator $\mathbf{S}_\alpha \colon \mathcal{H} \to L^2_\pi(\mathcal{X})$ as

$$\mathbf{S}_\alpha := \sum_{j\in N} \sigma_j^\alpha \ell_j \otimes h_j. \tag{47}$$

Clearly, we have that $\mathbf{S}_\pi = \mathbf{S}_1$, while $\mathrm{Im}\,\mathbf{S}_0 = \mathrm{cl}(\mathrm{Im}(S_\pi))$. Next, we equip $\mathrm{Im}(\mathbf{S}_\alpha)$ with a norm $\|\cdot\|_\alpha$ to build an interpolation space.

$$[\mathcal{H}]^c_\alpha := \left\{ f \in \mathrm{Im}(\mathbf{S}_\alpha) \mid \|f\|^2_\alpha := \sum_{j\in N} \sigma_j^{-2\alpha} \langle f, \ell_j\rangle^2 < \infty \right\}.$$

We remark that for $\alpha = 1$ the space $[\mathcal{H}]^c_\alpha$ is just an RKHS $\mathcal{H}$ seen as a subspace of $\mathrm{Im}(J_\pi) \subseteq L^2_\pi(\mathcal{X})$. Moreover, we have the following injections

$$[\mathcal{H}]^c_{\alpha_1} \hookrightarrow [\mathcal{H}]^c_1 \hookrightarrow [\mathcal{H}]^c_{\alpha_2} \hookrightarrow [\mathcal{H}]^c_0 \hookrightarrow \mathrm{Im}(J_\pi) \subseteq L^2_\pi(\mathcal{X}),$$

where $\alpha_1 \geq 1 \geq \alpha_2 \geq 0$.

Regularity condition. According to Zabczyk (2020, Theorem 2.2), the condition **(RC*)** is in fact equivalent to

$$\mathrm{Im}(\mathbf{A}_\pi \mathbf{S}_\pi) \subseteq \mathrm{Im}(\mathbf{S}_\alpha) \text{ and, hence, } \mathbf{A}_\pi \mathbf{S}_\pi = \mathbf{S}_\alpha \mathbf{G}^\alpha_{\mathcal{H}}, \text{ where } \mathbf{G}^\alpha_{\mathcal{H}} := \mathbf{S}^\dagger_\alpha \mathbf{T} \in \mathcal{B}(\mathcal{H}).$$

Remark B.1 (**Invariance of a RKHS**). When $\alpha \geq 1$, we necessarily have that $\mathrm{Im}(\mathbf{A}_\pi \mathbf{S}_\pi) \subseteq \mathrm{Im}(\mathbf{S}_\pi)$, i.e. $\mathcal{H}$ is π-a.e. invariant under the conditional expectation, and one has π-a.e. defined Koopman operator $\mathbf{G}_{\mathcal{H}} = \mathbf{G}^1_{\mathcal{H}}$. Moreover, since for $\alpha > 1$ we have that $\mathbf{G}_{\mathcal{H}} = \mathbf{C}^{\frac{\alpha-1}{2}} \mathbf{G}^\alpha_{\mathcal{H}}$, which implies that $\mathbf{G}_{\mathcal{H}}$ is compact, being product of a compact and bounded operators.

Embedding Property. Due to **(BK)** we also have that RKHS $\mathcal{H}$ can be embedded into $L^\infty_\pi(\mathcal{X})$, i.e. for some $\tau \in (0,1]$

$$[\mathcal{H}]^c_1 \hookrightarrow [\mathcal{H}]^c_\tau \hookrightarrow L^\infty_\pi(\mathcal{X}) \hookrightarrow L^2_\pi(\mathcal{X}),$$

Now, according to (Fischer & Steinwart, 2020), if $\mathbf{S}_{\tau,\infty} \colon [\mathcal{H}]^c_\tau \hookrightarrow L^\infty_\pi(\mathcal{X})$ denotes the injection operator, its boundedness implies the polynomial decay of the singular values of $\mathbf{S}_\pi$, i.e. $\sigma_j^2(\mathbf{S}_\pi) \lesssim j^{-1/\tau}$, $j \in N$, and the following condition is assured

(KE) *Kernel embedding property*: there exists $\tau \in [\beta, 1]$ such that

$$c_\tau := \|\mathbf{S}_{\tau,\infty}\|^2 = \operatorname*{ess\,sup}_{x\sim\pi} \sum_{j\in N} \sigma_j^{2\tau} |\ell_j(x)|^2 < +\infty. \tag{48}$$

Finally, we make the following remark on finite-dimensional RKHS.

Remark B.2 (**Finite-dimensional RKHS**). When $\mathcal{H}$ is finite dimensional, all spaces $[\mathcal{H}]_\alpha$ are finite dimensional. Hence, $\mathrm{Im}(\mathbf{A}_\pi \mathbf{S}_\pi) \subset \mathrm{Im}(\mathbf{S}_\pi)$ implies also $\mathrm{Im}(\mathbf{A}_\pi \mathbf{S}_\pi) \subset \mathrm{Im}(\mathbf{S}_\alpha)$ for every $\alpha > 0$. Moreover, we can set τ and β arbitrary close to zero.

Remark B.3 (**Link to CME**). Centering the feature map has been explored in the context of conditional mean embeddings (CME). The work most relevant to ours is (Klebanov et al., 2020), where one can find in-depth discussion on how kernel properties and centering affect the existence of $\mathbf{G}_{\mathcal{H}}$. On the other hand, the results in (Klebanov et al., 2020) are limited to the statistical consistency w.r.t. number of samples, while in this work we address finite sample learning rates and the impact of centering when learning dynamical systems.

B.2. Proof of Proposition 5.2

Embedding property and Whitened Feature Maps. The kernel embedding property **(KE)** allows one to estimate the norms of whitened centered feature maps $\xi(x) := \mathbf{C}_\gamma^{-1/2}[k_x - k_\pi]$, $\gamma > 0$, that play key role in deriving the learning rates, (Kostic et al., 2023a).

Proposition B.4. *Let* **(BK)** *and* **(SD*)** *hold for some* $\beta \in (0, 1]$, *and let* $\xi(x) := \boldsymbol{C}_\gamma^{-1/2}[k_x - k_\pi]$. *Then there exist* $\tau \in [\beta, 1]$ *and* $c_\tau, c_\beta \in (0, \infty)$ *such that for* $\gamma{>}0$

$$\|\xi\|^2 \leq c_\beta \gamma^{-\beta} \;\; \textit{and} \;\; \|\xi\|_\infty^2 \leq c_\tau \gamma^{-\tau}. \tag{24}$$

Proof. We first observe that for every $\tau > 0$ we have that

$$\begin{aligned}\|\xi(x)\|^2 &= \sum_{j\in N} \langle \mathbf{C}_\gamma^{-1/2}[k_x - k_\pi], h_j\rangle^2 = \sum_{j\in N} \frac{1}{\sigma_j^2+\gamma}\langle k_x - k_\pi, h_j\rangle^2 = \sum_{j\in N} \frac{\sigma_j^{2(1-\tau)}}{\sigma_j^2+\gamma}\frac{\langle k_x - k_\pi, h_j\rangle^2}{\sigma_j^2}\sigma_j^{2\tau} \\ &= \gamma^{-\tau}\sum_{j\in N}\frac{(\sigma_j^2\gamma^{-1})^{1-\tau}}{\sigma_j^2\gamma^{-1}+1}\frac{|h_j(x) - \mathbb{E}_{X\sim\pi}[h_j(X)]|^2}{\sigma_j^2}\sigma_j^{2\tau} \leq \gamma^{-\tau}\sum_{j\in N}\frac{|(\mathbf{S}_\pi h_j)(x)|^2}{\sigma_j^2}\sigma_j^{2\tau} = \gamma^{-\tau}\sum_{j\in N}|\ell_j(x)|^2\sigma_j^{2\tau},\end{aligned}$$

and, due to (48), we obtain $\|\xi\|_\infty^2 \leq \gamma^{-\tau}c_\tau$. On the other hand, we also have that

$$\|\xi\|^2 = \operatorname{tr}(\mathbb{E}_{X\sim\pi}[\xi(X)\otimes\xi(X)]) = \operatorname{tr}(\mathbf{C}_\gamma^{-1/2}\mathbf{C}\mathbf{C}_\gamma^{-1/2}) = \operatorname{tr}(\mathbf{C}_\gamma^{-1}\mathbf{C}),$$

which is in uncentered case known as effective dimension of the RKHS $\mathcal{H}$. Therefore, following the proof of Fischer & Steinwart (2020, Lemma 11) for uncentered covariances, we show that the bound on the effective dimension remains valid after centering with potentially improved bounds w.r.t. β. Namely, it holds that

$$\operatorname{tr}(\mathbf{C}_\gamma^{-1}\mathbf{C}) = \sum_{j\in N}\frac{\sigma_j^2}{\sigma_j^2+\gamma} \leq \begin{cases} \frac{b^\beta}{1-\beta}\gamma^{-\beta} & ,\beta < 1, \\ c_\tau\,\gamma^{-1} & ,\beta = 1. \end{cases} \tag{49}$$

For the case $\beta = 1$, it suffices to see that

$$\operatorname{tr}(\mathbf{C}_\gamma^{-1}\mathbf{C}) \leq \gamma^{-1}\sum_{j\in N}\sigma_j^2\|\ell_j\|^2 = \gamma^{-1}\int_{\mathcal{X}}\sum_{j\in N}\sigma_j^2|\ell_j(x)|^2\pi(dx) \leq \gamma^{-1}\operatorname*{ess\,sup}_{x\sim\pi}\sum_{j\in N}\sigma_j^2|\ell_j(x)|^2\pi(dx) = c_\tau\,\gamma^{-1},$$

while for $\beta < 1$ we can apply the same classical reasoning as in the proof of Proposition 3 of (Caponnetto & De Vito, 2007). □

B.3. Proof of Lemma 5.3

Lemma B.5. *Let* **(RC*)** *hold for some* $\alpha > 1$, *then there exists a compact* $\boldsymbol{G}_{\mathcal{H}}\colon \mathcal{H}\to\mathcal{H}$ *such that* $\boldsymbol{A}_\pi\boldsymbol{S}_\pi = \boldsymbol{S}_\pi\boldsymbol{G}_{\mathcal{H}}$ *and* $\rho(\boldsymbol{G}_{\mathcal{H}}){<}1$. *Consequently,* $\eta(\boldsymbol{G}_{\mathcal{H}}){<}\infty$, $d(\boldsymbol{G}_{\mathcal{H}}){>}0$ *and*

$$\frac{\eta(\boldsymbol{G}_{\mathcal{H}})d(\boldsymbol{G}_{\mathcal{H}})}{d(\boldsymbol{G}_{\mathcal{H}})+\|\boldsymbol{G}_{\mathcal{H}}-\widehat{\boldsymbol{G}}\|} \leq p(\widehat{\boldsymbol{G}}) \leq \frac{e}{2}\left[\frac{\eta(\boldsymbol{G}_{\mathcal{H}})d(\boldsymbol{G}_{\mathcal{H}})}{d(\boldsymbol{G}_{\mathcal{H}})-\|\boldsymbol{G}_{\mathcal{H}}-\widehat{\boldsymbol{G}}\|}\right]^2.$$

holds for every $\widehat{\boldsymbol{G}}$ *such that* $\|\boldsymbol{G}_{\mathcal{H}}-\widehat{\boldsymbol{G}}\| < d(\boldsymbol{G}_{\mathcal{H}})$.

Proof. First note that, according to Remark B.1, $\alpha > 1$ implies the existence of compact $\mathbf{G}_{\mathcal{H}}$, for which w.l.o.g. we can assume that $\mathbf{G}_{\mathcal{H}} \colon \operatorname{Ker}(\mathbf{S}_\pi)^\perp \to \operatorname{Ker}(\mathbf{S}_\pi)^\perp$. Since for $\mathbf{G}_{\mathcal{H}}$ the peripheral spectrum is a subset of the point spectrum, let λ be the leading eigenvalue of $\mathbf{G}_{\mathcal{H}}$ and by $h \in \mathcal{H} \setminus \{0\}$ its corresponding eigenvector. Then, since $\mathbf{A}_\pi \mathbf{S}_\pi = \mathbf{S}_\pi \mathbf{G}_{\mathcal{H}}$, we have that $\mathbf{A}_\pi \mathbf{S}_\pi h = \lambda \mathbf{S}_\pi h$ and, due to $\mathbf{S}_\pi h \neq 0$, we conclude that λ is an eigenvalue of $\mathbf{A}_\pi$, too. Therefore, $\rho(\mathbf{G}_{\mathcal{H}}) = |\lambda| \leq \rho(\mathbf{A}_\pi) < 1$.

Next, since $\eta(\widehat{\mathbf{G}}) \leq p(\widehat{\mathbf{G}}) \leq (e/2)[\eta(\widehat{\mathbf{G}})]^2$, it suffices to prove that for any two bounded operators A and Δ, one has that $|\eta(A) - \eta(A+\Delta)| \leq \eta(A)\|\Delta\|/(d(A) - \|\Delta\|)$.

To that end, denote $B = A + \Delta$ and observe that $(B - zI)^{-1} - (A - zI)^{-1} = (B - zI)^{-1}\Delta(A - zI)^{-1}$, and, hence,

$$|\|(B - zI)^{-1}\| - \|(A - zI)^{-1}\|| \leq \|(B - zI)^{-1}\|\|\Delta\|\|(A - zI)^{-1}\|,$$

i.e.

$$|\|(B - zI)^{-1}\|^{-1} - \|(A - zI)^{-1}\|^{-1}| \leq \|\Delta\|.$$

Now, recalling definition of the Kreiss constant we have that

$$\eta(B) = \sup_{|z|>1} \frac{|z|-1}{\|(B - zI)^{-1}\|^{-1}} \leq \sup_{|z|>1} \frac{|z|-1}{\|(A - zI)^{-1}\|^{-1} - \|\Delta\|} = \sup_{|z|>1} \frac{(|z|-1)\|(A - zI)^{-1}\|}{1 - \|\Delta\|/\|(A - zI)^{-1}\|^{-1}} \leq \frac{\eta(A)}{1 - \|\Delta\|/d(A)}.$$

Since, we can show the lower bound in an analogous way, the proof is completed. □

B.4. Proof of Theorems 5.4 and 6.1

We provide additional details used in the proof of this result.

Lemma B.6. *Let Assumption* **(RC*)** *be satisfied. Then*

$$\|\boldsymbol{G}_{\mathcal{H}} - \boldsymbol{G}_\gamma\|^2 \leq a^2 \gamma^{\alpha - 1}. \tag{50}$$

Proof of Lemma B.6. We have

$$\begin{aligned}\|\mathbf{G}_{\mathcal{H}} - \mathbf{G}_\gamma\|^2 &= \|\mathbf{C}_\gamma^{-1}\mathbf{T} - \mathbf{C}^\dagger \mathbf{T}\|^2 = \|(\mathbf{C}_\gamma^{-1} - \mathbf{C}^\dagger)\mathbf{T}\mathbf{T}^*(\mathbf{C}_\gamma^{-1} - \mathbf{C}^\dagger)\| \\ &\leq a^2 \|(\mathbf{C}_\gamma^{-1} - \mathbf{C}^\dagger)\mathbf{C}^{1+\alpha}(\mathbf{C}_\gamma^{-1} - \mathbf{C}^\dagger)\| = a^2 \gamma^{\alpha-1} \Big\| \sum_{j \,:\, \sigma_j > 0} \frac{(\gamma^{-1/2}\sigma_j)^{2(\alpha-1)}}{(1 + (\gamma^{-1/2}\sigma_j)^2)^2} h_j \otimes h_j \Big\|^2 \leq a^2 \gamma^{\alpha-1},\end{aligned}$$

where the last inequality holds due to $u^s \leq u + 1$ for all $u \geq 0$ and $s \in [0, 1]$ and using that the norm of the orthogonal projector $\sum_{j \,:\, \sigma_j > 0} h_j \otimes h_j$ equals one. □

To extend the proof of Theorem 5.4 to Theorem 6.1, it suffices to see that (29) and (30) imply that the forecasting error $\|\mu_t - \widehat{\mu}_t\|_{\mathcal{H}^*}$ can be upper bounded by $\|\mathbf{A}_\pi^t \mathbf{S}_\pi - \mathbf{S}_\pi \widehat{\mathbf{G}}^t\| \|q_0\| + \|\widehat{\mathbf{G}}^t\|(\|k_\pi - k_{\widehat{\pi}_x}\| + \|k_{\mu_0} - k_{\widehat{\mu}_0}\|) + \|k_\pi - k_{\widehat{\pi}_y}\|$. Hence, (9) applied to $\mathcal{E}_t^\circ$ and Briol et al. (2019, Lem. 1) that guarantees with probability at least $1 - \delta$ that $\|k_{\mu_0} - k_{\widehat{\mu}_0}\| \leq \epsilon$ with $\epsilon = \sqrt{\frac{2}{n_0} c_{\mathcal{H}}} \left(1 + \sqrt{\log(\delta^{-1})}\right)$, complete the proof.

B.5. Forecasting with RRR

We propose now to derive a result similar to Theorem 5.4 and 6.1 for the RRR estimator via an alternative argument which is valid for any $\alpha \in [1, 2]$ in (**(RC*)**). To that end, instead of Lemma 5.3 we will use the following result based on the Carleman-type bound on the resolvent of compact operators, see e.g. (Bandtlow, 2004).

Lemma B.7. *Let $\boldsymbol{G}$ be a bounded linear operator on a Hilbert space such that $\rho(\boldsymbol{G}) < 1$. If $\boldsymbol{G}$ has a finite rank r, then*

$$p(\boldsymbol{G}) := \sup_{t \in \mathbb{N}_0} \|\boldsymbol{G}^t\| \leq \frac{1}{2} \exp\left(\frac{2\, r\, \|\boldsymbol{G}\|}{1 - \rho(\boldsymbol{G})} + 1\right). \tag{51}$$

Proof. As before, we have that $p(\mathbf{G}) \leq (e/2)[\eta(\mathbf{G})]^2$, but now, since $\mathbf{G}$ is finite rank, we bound $\eta(\mathbf{G})$ using Carleman- type inequality. Namely, due to Bandtlow (2004, Theorem 4.1) for a trace-class operator A it holds that

$$\|(A - zI)^{-1}\| \leq \frac{1}{d(z, \mathrm{Sp}(A))} \exp\left(\frac{\|A\|_*}{d(z, \mathrm{Sp}(A))}\right),$$

where $d(z, \mathrm{Sp}(A)) := \min_{\omega \in \mathrm{Sp}(A)} |\omega - z|$ is the distance of $z \in \mathbb{C}$ to the spectrum of the operator A, and $\|\cdot\|_*$ denotes nuclear norm. Since, $\|\cdot\|_* \leq r \, \|A\|$ for A of finite rank r, using that $\mathrm{Sp}(\mathbf{G})$ is contained in the open unit disk, we obtain for $z \in \mathbb{C}$ s.t. $|z| > 1$

$$\|(z - \mathbf{G})^{-1}\|(|z| - 1) \leq \frac{(|z| - 1)}{d(z, \mathrm{Sp}(\mathbf{G}))} \exp\left(\frac{\|\mathbf{G}\|_*}{d(z, \mathrm{Sp}(\mathbf{G}))}\right) \leq \exp\left(\frac{r\|\mathbf{G}\|}{d(z, \mathrm{Sp}(\mathbf{G}))}\right),$$

and, thus,

$$\eta(\mathbf{G}) \leq \exp\left(\frac{r\|\mathbf{G}\|}{\inf_{|z|>1} d(z, \mathrm{Sp}(\mathbf{G}))}\right) \leq \exp\left(\frac{r\|\mathbf{G}\|}{1 - \rho(\mathbf{G})}\right).$$

□

Corollary B.8. *Assume the operator $\boldsymbol{A}_\pi$ is of finite rank r for some $r \in \mathbb{N}$. Let* **(SD*)** *and* **(RC*)** *hold for some $\beta \in (0, 1]$ and $\alpha \in [1, 2]$, respectively. In addition, let* $\mathrm{cl}(\mathrm{Im}(\boldsymbol{S}_\pi)) = L^2_\pi(\mathcal{X})$ *and* **(BK)** *be satisfied. Let*

$$\gamma \asymp n^{-\frac{1}{\alpha+\beta}} \text{ and } \varepsilon_n^\star := n^{-\frac{\alpha}{2(\alpha+\beta)}}.$$

Let $\delta \in (0, 1)$. Then the forecasted distributions (31) based on $\widehat{\boldsymbol{G}} = \widehat{\boldsymbol{G}}^{\mathrm{RRR}}_{r,\gamma}$ satisfy for n large enough, with probability at least $1 - \delta$, for any $t \geq 1$

$$\|\widehat{\mu}_t - \mu_t\|_{\mathcal{H}^*} \lesssim_{c_\mathcal{H}} e^{\frac{8r}{1-\rho(\boldsymbol{A}_\pi)}} \left(\left(a + \frac{1}{\sigma_r^2(\boldsymbol{A}_\pi \boldsymbol{S}_\pi)}\right) \varepsilon_n^\star \, \ln(\delta^{-1}) + \sqrt{\frac{\ln \delta^{-1}}{n_0 \wedge n}}\right).$$

Proof. For brevity, we set $\widehat{\mathbf{G}} = \widehat{\mathbf{G}}^{\mathrm{RRR}}_{r,\gamma}$ and $\mathbf{G} = \mathbf{G}^{\mathrm{RRR}}_{r,\gamma}$. Exploiting the definition and properties of the RRR model, we prove that there exists a constant $c > 0$, depending only $r, c_\mathcal{H}, \beta$ such that for large enough $n \geq r$, with probability at least $1 - \delta$, the estimator $\widehat{\mathbf{G}} = \widehat{\mathbf{G}}^{\mathrm{RRR}}_{r,\gamma}$ satisfies $\|\widehat{\mathbf{G}}\| \leq 2$, $1 - \rho(\widehat{\mathbf{G}}) \geq \frac{1-\rho(\mathbf{G})}{2}$ and $\mathcal{E}^\circ(\widehat{\mathbf{G}}) \lesssim_{c_\mathcal{H}} \varepsilon_n^\star \, \ln(\delta^{-1})$.

We first observed that

$$\mathcal{E}^\circ(\widehat{\mathbf{G}}) \leq \|\mathbf{A}_\pi \mathbf{S}_\pi - \mathbf{S}_\pi \mathbf{G}_\gamma\| + \|\mathbf{S}_\pi(\mathbf{G}_\gamma - \mathbf{G})\| + \|\mathbf{S}_\pi(\widehat{\mathbf{G}} - \mathbf{G})\|.$$

Proposition 5 in (Kostic et al., 2023a) and the condition $\mathrm{cl}(\mathrm{Im}(\mathbf{S}_\pi)) = L^2_\pi(\mathcal{X})$ immediately give $\|\mathbf{A}_\pi \mathbf{S}_\pi - \mathbf{S}_\pi \mathbf{G}\| \leq a\gamma^{\alpha/2}$, since for universal kernel, we have $\mathrm{Im}(\mathbf{A}_\pi \mathbf{S}_\pi) \subseteq \mathrm{cl}(\mathrm{Im}(\mathbf{S}_\pi))$. By definition of $\mathbf{G}_\gamma$ and since $\mathrm{rank}(\mathbf{A}_\pi) = r$, we have $\mathbf{G}^{\mathrm{RRR}}_{r,\gamma} = \mathbf{G}_\gamma$. Hence $\|\mathbf{S}_\pi(\mathbf{G}_\gamma - \mathbf{G})\| = 0$.

We prove below that there exists a constant $c = c(c_\mathcal{H}) > 0$ such that, for $n \geq r$ large enough

$$\mathbb{P}\left\{\|\mathbf{S}_\pi(\widehat{\mathbf{G}} - \mathbf{G})\| \leq c \, r^{\frac{2}{\beta}} \sqrt{\frac{1}{n\gamma^\beta}} \ln(\delta^{-1})\right\} \geq 1 - \delta. \tag{52}$$

Combining the previous display with our control on the bias and using that $\gamma \asymp n^{-\frac{1}{\alpha+\beta}}$, we get that

$$\mathbb{P}\left\{\mathcal{E}^\circ(\widehat{\mathbf{G}}) \lesssim_{c_\mathcal{H}} \left(a + r^{\frac{2}{\beta}}\right) n^{-\frac{\alpha}{2(\alpha+\beta)}} \log(\delta^{-1})\right\} \geq 1 - \delta.$$

Define $\widehat{B} := \widehat{\mathbf{C}}_\gamma^{-1/2} \widehat{\mathbf{T}}$ and let $\widehat{P}_r$ denote the orthogonal projector onto the subspace of leading r right singular vectors of $\widehat{B}$. Then we have $\widehat{\mathbf{G}}^{\mathrm{RRR}}_{r,\gamma} = \widehat{G}_\gamma \widehat{P}_r$. Hence, we have $\|\widehat{\mathbf{G}}^{\mathrm{RRR}}_{r,\gamma}\| \leq \|\widehat{\mathbf{G}}_\gamma\|$. Exploiting Proposition 16 in (Kostic et al., 2023a), we prove below that for n large enough

$$\mathbb{P}\left\{\|\widehat{\mathbf{G}}\| \leq 2\right\} \geq 1 - \delta. \tag{53}$$

Next we apply Corollary 1 of (Kostic et al., 2023a) to obtain that

$$\mathbb{P}\left\{\rho(\widehat{\mathbf{G}}) \leq \rho(\mathbf{A}_\pi) + \varepsilon_n^\star \ln(\delta^{-1})\right\} \geq 1 - \delta. \tag{54}$$

Consequently on the same event, provided that n is large enough, we deduce that

$$1 - \rho(\widehat{\mathbf{G}}) \geq \frac{1 - \rho(\mathbf{A}_\pi)}{2}.$$

An elementary union bound combining the previous results and Lemma B.7 below gives the result with probability at least $1 - 5\delta$. Up to a rescaling of the constant, we can replace $1 - 5\delta$ by $1 - \delta$.

□

Proof of Equation (52). We first define $B := \mathbf{C}_\gamma^{-1/2}\mathbf{T}$ and we recall that $\mathbf{T} = \mathbf{S}_\pi^* \mathbf{A}_\pi \mathbf{S}_\pi$. Applying Proposition 18 in (Kostic et al., 2023a) gives, with probability at least $1 - \delta$,

$$\|\mathbf{S}_\pi(\mathbf{G}_{r,\gamma}^{\mathrm{RRR}} - \widehat{\mathbf{G}}_{r,\gamma}^{\mathrm{RRR}})\| \leq \frac{c\,\varepsilon_n^2(\gamma,\delta/5)}{1 - \varepsilon_n^1(\gamma,\delta/5)} + \frac{\sigma_1(B)}{\sigma_r^2(B) - \sigma_{r+1}^2(B)} \frac{(c^2-1)\,\varepsilon_n(\delta/5) + c^2\,(\varepsilon_n^2(\gamma,\delta/5))^2}{(1 - \varepsilon_n^1(\gamma,\delta/5))^2}, \tag{55}$$

where $c := 1 + a\, c_{\mathcal{H}}^{(\alpha-1)/2}$,

$$\varepsilon_n(\delta) := \frac{4c_{\mathcal{H}}}{3n}\mathcal{L}(\delta) + \sqrt{\frac{2\|\mathbf{C}\|}{n}\mathcal{L}(\delta)} \quad \text{and} \quad \mathcal{L}(\delta) := \log\frac{4\,\mathrm{tr}(\mathbf{C})}{\delta\,\|\mathbf{C}\|}, \tag{56}$$

$$\varepsilon_n^1(\gamma,\delta) := \frac{4c_\tau}{3n\gamma^\tau}\mathcal{L}^1(\gamma,\delta) + \sqrt{\frac{2\,c_\tau}{n\,\gamma^\tau}\mathcal{L}^1(\gamma,\delta)}, \tag{57}$$

with

$$\mathcal{L}^1(\gamma,\delta) := \log\frac{4}{\delta} + \log\frac{\mathrm{tr}(\mathbf{C}_\gamma^{-1}\mathbf{C})}{\|\mathbf{C}_\gamma^{-1}\mathbf{C}\|},$$

and

$$\varepsilon_n^2(\gamma,\delta) := 4\sqrt{2\,c_{\mathcal{H}}}\left(\sqrt{\frac{\mathrm{tr}(\mathbf{C}_\gamma^{-1}\mathbf{C})}{n}} + \frac{\sqrt{c_\tau}}{n\gamma^{\tau/2}}\right)\log\frac{2}{\delta}. \tag{58}$$

Using (49) and elementary computations, we get that the dominating term in (55) is of the order $\sqrt{\frac{1}{n\gamma^\beta}}\ln(\delta^{-1})$ since $\tau \geq \beta$. In addition, for $B := \mathbf{C}_\gamma^{-1/2}\mathbf{T}$, we have $\sigma_{r+1}(B) = 0$ since $\mathrm{rank}(\mathbf{T}) \leq \mathrm{rank}(\mathbf{A}_\pi) = 5$. Finally, Proposition 6 of (Kostic et al., 2023a) and our choice of γ guarantees for n large enough that $\sigma_r^2(B) \geq \sigma_r^2(\mathbf{A}_\pi\mathbf{S}_\pi) - a^2\, c_{\mathcal{H}}^{\alpha/2}\,\gamma^{\alpha/2} \geq \sigma_r^2(\mathbf{A}_\pi\mathbf{S}_\pi)/2 > 0$.

Proof of Equation (53). Proposition 16 in (Kostic et al., 2023a) guarantees with probability at least $1 - \delta$

$$\|\widehat{\mathbf{G}}_\gamma\| \leq \frac{1 + \varepsilon_n^3(\gamma,\delta/2)}{1 - \varepsilon_n^3(\gamma,\delta/2)},$$

where

$$\varepsilon_n^3(\gamma,\delta) := 4\sqrt{2\,c_{\mathcal{H}}}\left(\sqrt{\frac{\mathrm{tr}(\mathbf{C}_\gamma^{-2}\mathbf{C})}{n}} + \frac{\sqrt{c_\tau}}{n\gamma^{(1+\tau)/2}}\right)\log\frac{2}{\delta}. \tag{59}$$

Using again (49) and the fact that $\tau \geq \beta$, we deduce that

$$\varepsilon_n^3(\gamma,\delta) \lesssim \sqrt{c_{\mathcal{H}}}\sqrt{\frac{1}{n\gamma^{1+\beta}}}\log(2\delta^{-1}).$$

With our choice of γ and for n large enough such that $\varepsilon_n^3(\gamma,\delta/2) < 1/4$, we get

$$\mathbb{P}\left\{\|\widehat{\mathbf{G}}_\gamma\| \leq \frac{1 + \varepsilon_n^3(\gamma,\delta/2)}{1 - \varepsilon_n^3(\gamma,\delta/2)} \leq 2\right\} \geq 1 - \delta.$$

B.6. Forecasting the Conditional Variance

In this section we focus on the conditional variance of an observable, and how to estimate it using DLI framework. First, note that fixing the observable $h \in \mathcal{H}$ and time-step $t \in \mathbb{N}$ and considering joint distribution of (X_0, X_t), according to standard bias-variance decomposition of the square regression loss, any learner $\widehat{h}_t$ satisfies

$$\underbrace{\mathbb{E}_{(X_0,X_t)}[h(X_t) - \widehat{h}_t(X_0))]^2}_{MSE} = \underbrace{\mathbb{E}_{X_0}[\mathbb{E}[h(X_t) \,|\, X_0] - \widehat{h}_t(X_0)]^2}_{\text{bias}} + \underbrace{\mathbb{E}_{X_t}[h(X_t)]^2 - \mathbb{E}_{X_0}[\mathbb{E}[h(X_t) \,|\, X_0]]^2}_{\text{variance}}. \tag{60}$$

Hence, the irreducible part of the mean square error is the property of the distribution of the data and the observable.

Using the tower property of the expectation, we can further express the term $\mathbb{E}_{X_0}\mathbb{V}[h(X_t) \,|\, X_0]$, where the conditional variance is defined as

$$\mathbb{V}[h(X_t) \,|\, X_0] := \mathbb{E}[[h(X_t)]^2 \,|\, X_0] - [\mathbb{E}[h(X_t) \,|\, X_0]]^2. \tag{61}$$

Now, assuming that both functions h and $[h(\cdot)]^2$ belong to the RKHS $\mathcal{H}$, we can write the conditional variance via Koopman operators in order to estiamte is

$$\mathbb{V}[h(X_t) \,|\, X_0 = x] := [A_\pi^t S_\pi [h(\cdot)]^2](x) - [[A_\pi^t S_\pi h](x)]^2 \approx \widehat{h(\cdot)^2}_t(x) - (\widehat{h}_t(x))^2 =: \widetilde{h}_t(x), \tag{62}$$

which leads to the obvious estimator. But then, the following lemma allows us to extend the bounds on the $L^2_\pi(\mathcal{X})$ estimation error of conditional mean to the to the $L^1_\pi(\mathcal{X})$ estimation error of the conditional variance. That is, we have the following result matching the one of Theorem 5.4, whose proof is a direct consequence of the subsequent lemma.

Theorem B.9. *Let* **(SD*)** *and* **(RC*)** *hold for some $\beta \in (0, 1]$ and $\alpha \in (1, 2]$, respectively. In addition, let* $\mathrm{cl}(\mathrm{Im}(S_\pi)) = L^2_\pi(\mathcal{X})$ *and* **(BK)** *be satisfied. If $\delta \in (0, 1)$,*

$$\gamma \asymp n^{-\frac{1}{\alpha+\beta}} \;\text{ and }\; \varepsilon_n^\star := n^{-\frac{\alpha}{2(\alpha+\beta)}},$$

then, for every $t \in \mathbb{N}$, the forecasted conditional variance of the observable given in (62) based on KRR satisfies

$$\|\mathbb{V}[h(X_t) \,|\, X_0 = \cdot] - \widetilde{h}_t\|_{L^1_\pi(\mathcal{X})} / \|h\|^2 \leq C\varepsilon_n^\star \,\ln(\delta^{-1}),$$

with probability at least $1 - \delta$ w.r.t. iid sampled data $\mathcal{D}$ according to the invariant distribution π, where the constant C may depend only on a, b and $c_\mathcal{H}$.

Lemma B.10. *Given $x \in \mathcal{X}$ and $t \in \mathbb{N}$, let $E_t \colon \mathcal{H} \to L^2_\pi(\mathcal{X})$ be a bounded linear operator, and define a (non-linear) operator $V_t \colon \mathcal{H} \to L^1_\pi(\mathcal{X})$ as $[V_t h](x) := [E_t(h(\cdot)^2)](x) - [E_t h](x)^2$, defined on $\{h \in \mathcal{H} \,|\, h(\cdot)^2 \in \mathcal{H}\}$. Then for $\varepsilon > 0$ the following holds*

$$\|\mathbb{E}[h(X_t) \,|\, X_0 = \cdot] - E_t h\|_{L^2_\pi(\mathcal{X})} / \|h\|_\mathcal{H} \leq \varepsilon \implies \|\mathbb{V}[h(X_t) \,|\, X_0 = \cdot] - V_t h\|_{L^1_\pi(\mathcal{X})} \leq \varepsilon(2\sqrt{c_\mathcal{H}} + \varepsilon)\|h\|^2_\mathcal{H}.$$

Proof. From the definition of V_t and (62), we have that

$$\|\mathbb{V}[h(X_t) \,|\, X_0 = \cdot] - V_t h\|_{L^1_\pi(\mathcal{X})} \leq \|\mathbb{E}[g(X_t) \,|\, X_0 = \cdot] - E_t g\|_{L^1_\pi(\mathcal{X})} + \mathbb{E}_{x\sim\pi}|\mathbb{E}[h(X_t) \,|\, X_0 = x]^2 - [E_t h](x)^2|.$$

Since $L^1_\pi(\mathcal{X})$ norm is bounded by $L^2_\pi(\mathcal{X})$ norm and $\|g\| \leq \|h\|^2$, the first term is bounded by $\|h\|^2\varepsilon$.

For the second term, observe that

$$\mathbb{E}_{x\sim\pi}|\mathbb{E}[h(X_t) \,|\, X_0 = x]^2 - [E_t h](x)^2| = \int_\mathcal{X} |\mathbb{E}[h(X_t) \,|\, X_0 = x] - [E_t h](x)||\mathbb{E}[h(X_t) \,|\, X_0 = x] + [E_t h](x)|\pi(dx).$$

Hence, using the Cauchy-Schwartz inequality, we can bound it by

$$\|\mathbb{E}[h(X_t) \,|\, X_0 = \cdot] - E_t h\|_{L^2_\pi(\mathcal{X})}\|\mathbb{E}[h(X_t) \,|\, X_0 = \cdot] + E_t h\|_{L^2_\pi(\mathcal{X})},$$

and, consequently, by $\varepsilon\|h\|\,(2\|A_\pi^t S_\pi h\| + \varepsilon\|h\|)$, which yields the bound $\varepsilon(2\sqrt{c_\mathcal{H}} + \varepsilon)\|h\|^2$ and completes the proof. □

Therefore, the high probability forecasting error bounds that hold for conditional mean DLI estimators, hold also for the corresponding conditional variance estimators in en adequate norm. Thus, theorems on RRR and PCR estimators are readily extended to cover the conditional variance estimation.

B.7. Non-iid Samples from a Trajectory

In this section we show how that the results of this paper based on iid assumption of the data samples are seamlessly extended to the case of sampling along the trajectory. More precisely, we consider that a trajectory $x_1, \dots, x_{n+1}$ has been sampled from the process as $x_1 \sim \pi, y_{k-1} = x_k \sim p(x_{k-1}, \cdot)$, $k \in [2{:}n]$, and rely on the basic strategy, going back to at least (Yu, 1994), that represents the process $(X_t)_{t\in\mathbb{N}}$ by two interlaced block-processes in order to transfer a concentration result for i.i.d. variables to the non-i.i.d. case. Such block-processes $(Y_t)_{t\in\mathbb{N}}$ and $(Y'_t)_{t\in\mathbb{N}}$ are defined as

$$Y_t = \sum_{k=2(t-1)s+1}^{(2t-1)s} X_k \quad \text{and} \quad Y'_t = \sum_{k=(2t-1)s+1}^{2ts} X_k \quad \text{for } t \in \mathbb{N},$$

to accomplish that Y_t/Y'_t and Y_{t+1}/Y'_{t+1} are sufficiently separated to be regarded as independent.

This is possible assuming that the Markov process is β-mixing, that is for $\tau \in \mathbb{N}$ we can define coefficients $\beta_p(\tau)$ as

$$\beta_p(s) = \sup_{B\in\Sigma\otimes\Sigma} |\rho_s(B) - (\pi\times\pi)(B)|,$$

where ρ_s is the joint distribution of X_1 and X_{1+s}, that tend to zero with increasing $s \in \mathbb{N}$. Then, our extension to non-iid setting is based on the following Lemma 1 of Kostic et al. (2022), which we restate here.

Lemma B.11. *Let $(X_t)_{t\in\mathbb{N}}$ be strictly stationary with values in a normed space $(\mathcal{X}, \|\cdot\|)$, and assume $n = 2ms$ for $s, m \in \mathbb{N}$. Moreover, let $Z_1, \dots, Z_m$ be m independent copies of $Z_1 = \sum_{k=1}^{s} X_k$. Then for $\varepsilon > 0$*

$$\mathbb{P}\Big\{\Big\|\sum_{i=1}^{n} X_i\Big\| > \varepsilon\Big\} \le 2\,\mathbb{P}\Big\{\Big\|\sum_{j=1}^{m} Z_j\Big\| > \frac{\varepsilon}{2}\Big\} + 2\,(m-1)\,\beta_p(s).$$

As an application of this result we can transfer the centered versions of the Propositions 12-15 from Kostic et al. (2023a) which were proved in the i.i.d. setting to the non-iid setting. For brevity we showcase how this is done on Kostic et al. (2023a, Proposition 12), while the rest follows in an analogous way using whitened features.

Proposition B.12. *Let $\delta > 2(m-1)\beta_p(s)$. With probability at least $1-\delta$ in the draw $x_1 \sim \pi, x_i \sim p(x_{i-1}, \cdot)$, $i \in [2{:}n]$, it holds that*

$$\mathbb{P}\{\|\widehat{\boldsymbol{T}} - \boldsymbol{T}\| \le \varepsilon_n(\delta)\} \wedge \mathbb{P}\{\|\widehat{\boldsymbol{C}} - \boldsymbol{C}\| \le \varepsilon_n(\delta)\} \ge 1-\delta,$$

where

$$\varepsilon_n(\delta) := \frac{4c_{\mathcal{H}}}{3(n/2s)}\mathcal{L}(\delta) + \sqrt{\frac{2m\|C\|}{(n/2s)^2}\mathcal{L}(\delta)} \quad \textit{and} \quad \mathcal{L}(\delta) := \log\frac{4\operatorname{tr}(C)}{(\delta/2-(m-1)\beta_p(s))\,\|C\|}. \tag{63}$$

Proof. Given $s \in \mathbb{N}$, let $Z_1, \dots, Z_m$ be independent copies of $Z_1 = \sum_{i=1}^{s}[\phi(x_i) - k_\pi] \otimes [\phi(x_{i+1}) - k_\pi] - s\mathbf{T}$. Now, applying Lemma B.11 with $[\phi(x_i) - k_\pi] \otimes [\phi(x_{i+1}) - k_\pi] - \mathbf{T}$ in place of X_i we obtain

$$\mathbb{P}\left\{\left\|\widehat{\mathbf{T}} - \mathbf{T}\right\| > \varepsilon\right\} = \mathbb{P}\left\{\left\|\sum_{i=1}^{n}[\phi(x_i) - k_\pi] \otimes [\phi(x_{i+1}) - k_\pi] - \mathbf{T}\right\| > n\varepsilon\right\} \le 2\mathbb{P}\left\{\left\|\sum_{j=1}^{m} Z_j\right\| > \frac{n\varepsilon}{2}\right\} + 2\,(m-1)\,\beta_p(s).$$

To obtain the result whp probability $1-\delta$, we bound the rightmost probability using the non-commutative Bernstein inequality of Kostic et al. (2023a, Proposition 11) with iid operators Z_i setting the probability as $\delta/2 - (m-1)\beta_p(s)$ to obtain

$$\mathbb{P}\left\{\left\|\widehat{\mathbf{T}} - \mathbf{T}\right\| > \varepsilon\right\} \le 2\mathbb{P}\left\{\left\|\frac{1}{m}\sum_{j=1}^{m} Z_j\right\| > \frac{n\varepsilon}{2ms}\right\} + 2\,(m-1)\,\beta_p(s) \le \delta.$$

Finally, solving for ε we obtain the proof for the centered cross-covariance. In the same way we obtain the bound for the covariance. □

We notice that this result, apart from slightly larger numerical constants in a logarithmic term bound (63), is conceptually identical to Kostic et al. (2023a, Proposition 12), when the sample size n is replaced by the *effective sample size* $m \approx n/2s$. The same conclusion remains true for Propositions 13-15 of Kostic et al. (2023a), since the only difference lies in the impact of the regularization parameter $\gamma > 0$ on the bound. Therefore, we conclude that when the sampling of the data $\mathcal{D}_n$ is done from a trajectory, in the results of Theorems 5.4 and 6.1 we have the bound $(n/2s)^{-\frac{\alpha}{2(\alpha+\beta)}} \log \frac{2}{\delta}$, where $s \in \mathbb{N}$ is such that $(n/2s-1)\beta_p(s) \leq \delta/2$.

C. Experimental Details

In both experiments, the Reduced Rank Regression estimator was implemented using the reference code from (Kostic et al., 2022) available at https://github.com/Machine-Learning-Dynamical-Systems/kooplearn. The experiments were run on a workstation equipped with an Intel(R) Core™i9-9900X CPU @ 3.50GHz, 48GB of RAM and a NVIDIA GeForce RTX 2080 Ti GPU. All experiments have been implemented in Python 3.11.

For the Cox–Ingersoll–Ross, the conditional expectation of the state and its variance are given by

$$\mathbb{E}\left[r_t \mid r_0 = r\right] = re^{-at} + b(1 - e^{-at})$$

and

$$\mathbb{V}\left[r_t \mid r_0 = r\right] = r\frac{\sigma^2}{a}(e^{-at} - e^{-2at}) + \frac{b\sigma^2}{2a}(1-e^{-at})^2.$$

For the *Ornstein-Uhlenbeck* experiment we sampled the process every $dt = 0.05$. The estimators were trained with 250 observations sampled independently from the invariant distribution, while the initial distribution used to evaluate the MMD was sampled 1000 times. Each experiment has been repeated 100 times independently, and the hyperparameters were tuned on a validation set of 500 points sampled from the invariant distribution.

The code to reproduce the experiments will be open sourced.