Scalable Infinitesimal Generator–Based Koopman Learning for Long-Horizon Prediction

Minseok Jeong[1] DSA950115@KAIST.AC.KR
SooJean Han[1] SOOJEAN@KAIST.AC.KR
Hyo-Sang Shin[2] HYOSANGSHIN@KAIST.AC.KR
[1] *School of Electrical Engineering, KAIST, Daejeon, Republic of Korea*
[2] *Cho Chun Shik Graduate School of Mobility, KAIST, Daejeon, Republic of Korea*

Editors: G. Sukhatme, L. Lindemann, S. Tu, A. Wierman, N. Atanasov

Abstract

Koopman operator theory offers a linear representation of nonlinear dynamics and has strong potential for long-horizon prediction. However, most existing methods rely on one-step snapshot fitting and suffer from error accumulation over long rollouts. Prior work has attempted to address this by extending training horizons or using physics-informed, generator-based formulations. Still, these approaches remain limited, partly because they rely on MLP-based observables, whose spectral bias favors low-frequency components.

In this paper, we introduce a Random Fourier Feature–lifted physics-informed Koopman network (RFF-PIKN) that directly minimizes the generator loss. We first show that snapshot-based local transition fitting has inherent limitations in long-horizon stability relative to generator-based learning. We then prove that RFF-PIKN converges stably to a local minimizer of the true population risk under the generator loss. Finally, empirical comparisons demonstrate that RFF-PIKN outperforms MLP- and polynomial-based observables in long-horizon prediction while substantially reducing computation, and further matches key behaviors of oracle kernel-based methods.

Keywords: Koopman operator learning, Physics-informed learning, Long-horizon prediction, Reproducing kernel Hilbert space

1. Introduction

Koopman operator theory provides a linear representation of nonlinear dynamical systems. Once the operator is identified, trajectories can be propagated through repeated matrix multiplications, offering a more efficient alternative to direct simulation of nonlinear dynamics (Mezić, 2005; Klus et al., 2020b). This advantage makes Koopman-based approaches attractive in applications where long-horizon rollout is essential, such as stability analysis in power systems (Korda and Mezić, 2018), trajectory planning in robotics and autonomous systems (Brunton et al., 2016; Williams et al., 2015), and reachability analysis (Budišić et al., 2012).

Limitations of Snapshot-based Schemes. Many current methods, particularly the widely used snapshot-based formulations such as extended dynamic mode decomposition (EDMD) (Williams et al., 2015; Tu et al., 2014), struggle to produce reliable long-horizon rollouts. By design, these approaches optimize for one-step prediction accuracy, capturing local transitions but neglecting consistency throughout the global flow of the dynamics. As a result, errors accumulate rapidly over extended horizons, and prediction quality deteriorates (Mezić, 2005; Brunton et al., 2016). We emphasize that this short-term bias is not merely an empirical artifact but a structural limitation of snapshot-based losses. Without additional assumptions such as ergodicity and highly expressive feature maps, these methods are fundamentally ill-suited for long-horizon prediction.

The Role of Generator-based Formulations. A more principled way to address long-horizon degradation is to move from snapshot-based losses to generator-based formulations. By constraining the infinitesimal generator, these approaches enforce local consistency with the underlying vector field, thereby curbing error accumulation over extended horizons. Recent advances in physics-informed learning (Raissi and Karniadakis, 2017; Raissi et al., 2019; Karniadakis et al., 2021) and their extensions to Koopman frameworks (Klus et al., 2020b; Liu et al., 2024) highlight this potential. In particular, the physics-informed Koopman network (PIKN) (Liu et al., 2024) enforces generator-level consistency by embedding model knowledge directly into Koopman learning.

Remaining Challenges. Despite the growing availability of model-based environments with known dynamics $\mathbf{f}$, research explicitly targeting model-based long-horizon prediction remains limited (Gürtler and Martius, 2025; Castiglione et al., 2024). Most existing approaches simply extend snapshot-based schemes with minor modifications, such as adding long-horizon error regularization (Xiao et al., 2019; Sam et al., 2023). However, they fail to fully exploit the advantages of model-based environments, including unlimited resampling, scalable feature representations, and real-time updates.

As briefly examined in the PIKN paper (Liu et al., 2024), the infinitesimal generator–based approach shows promise for reliable long-horizon prediction in model-based settings. However, its theoretical analysis remains underexplored, leaving open *why generator-level consistency should guarantee superior long-horizon behavior*. Moreover, even when effective in principle, generator-based approaches such as PIKN face practical challenges. The MLP-based observables used in PIKN suffer from spectral bias (Rahaman et al., 2019), limiting their ability to capture high-frequency components. In addition, computing Jacobians of feature mappings is computationally expensive, which undermines scalability.

Contributions. In this work, we address these issues by systematically examining the limitations of snapshot-based Koopman learning for long-horizon prediction and proposing RFF-PIKN, an approach to mitigate them. The key contributions of this work are as follows:

- **Analytic Characterization of Snapshot Losses.** We provide theoretical analyses showing that snapshot-based losses are fundamentally limited for long-horizon prediction. Although they can capture short-term dynamics, their dependence on discretization methods that underutilize physical consistency misaligns them with the true flow. Motivated by this limitation, we advocate for a generator-based learning scheme.

- **Novel Koopman Learning Framework.** We employ a Random Fourier Feature (RFF) approximation in the observable function space. This choice mitigates the spectral bias inherent in MLP-based observables. Additionally, its lightweight structure enables efficient online approximation of kernel-based methods. In turn, this property supports scalable training in model-based settings with frequent resampling.

- **Convergence and Learning Behavior of Estimated Parameters.** We establish convergence guarantees for the proposed framework by interpreting resampling as stochastic batch sampling in stochastic gradient descent (SGD). We further compare the properties of the estimators with a kernel-based method regarded as the ground truth. This analysis provides a systematic understanding of the learning dynamics of both the observable and operator.

2. Preliminaries

2.1. Notations

We consider an autonomous continuous-time nonlinear dynamical system $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}),\ \mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d$, where $\mathbf{f}$ denotes the vector field governing the dynamics. Let $C^p(\mathcal{X}; \mathbb{R}^d)$ denote the space of p-times continuously differentiable functions mapping from $\mathcal{X}$ to $\mathbb{R}^d$. We denote by $\mathbf{F}^t$ the time-t flow map of the system, which satisfies $\mathbf{F}^{t+s}(\mathbf{x}) = \mathbf{F}^t(\mathbf{F}^s(\mathbf{x}))$ for all $\mathbf{x} \in \mathcal{X}$ and $t, s \geq 0$.

We denote by $\mathcal{G} = \{g : \mathcal{X} \to \mathbb{R}\}$ the space of observables defined on $\mathcal{X}$. Furthermore, we define a function class $\widehat{\mathcal{G}} = \{g : \mathcal{X} \to \mathbb{R} \mid g \text{ satisfies prescribed structural assumptions}\}$, which approximates either a finite-dimensional subspace (e.g., polynomial basis) or the inner-product structure of $\mathcal{G}$ (e.g., RKHS).

The infinitesimal generator of the Koopman operator is defined by $\mathbb{L}g = \lim_{t\to 0}(g \circ \mathbf{F}^t - g)/t$, for $g \in \mathcal{G}$. Its discrete-time counterpart over a time step Δt is given by $\mathbb{A}_{\Delta t} = e^{\mathbb{L}\Delta t}$. For notational simplicity, we will write $\mathbb{A} \equiv \mathbb{A}_{\Delta t}$ and $\widetilde{\mathbf{F}}(\mathbf{x}) \equiv \mathbf{F}^{\Delta t}(\mathbf{x})$. Their finite-dimensional matrix representations are denoted by $\mathbf{L}$ and $\mathbf{A} \equiv \mathbf{A}_{\Delta t}$, respectively. To avoid confusion with the kernel matrix $\mathbf{K}$, which will appear later in RKHS-based formulations, we consistently use $\mathbf{A}$ to denote the finite-dimensional Koopman operator.

We denote a D-dimensional observable map by $\mathbf{g} = (g_1, \ldots, g_D) : \mathcal{X} \to \mathbb{R}^D$, with each $g_i \in \mathcal{G}$ or $\widehat{\mathcal{G}}$. For a vector $\mathbf{g}(\mathbf{x}) \in \mathbb{R}^D$, $\|\mathbf{g}(\mathbf{x})\|_p$ denotes its ℓ^p norm. On a probability space $(\mathcal{X}, \mathcal{B}, \nu)$, we denote expectation by $\mathbb{E}_{\mathbf{x}\sim\nu}$ and abbreviate it as $\mathbb{E}_\nu$. The L^p space is $L^p(\nu; \mathbb{R}^D) = \{\mathbf{g} : \mathcal{X} \to \mathbb{R}^D \mid \mathbb{E}_\nu[\|\mathbf{g}(\mathbf{x})\|_2^p] < \infty\}$. In particular, $L^2(\nu; \mathbb{R}^D)$ is a Hilbert space with inner product $\langle \mathbf{g}, \mathbf{h} \rangle = \mathbb{E}_\nu[\mathbf{g}(\mathbf{x})^\top \mathbf{h}(\mathbf{x})]$, which we abbreviate as $\mathbb{E}_\nu[\mathbf{g}^\top \mathbf{h}]$ when the dependence on $\mathbf{x}$ is clear.

2.2. Distinct Learning Frameworks

Throughout the paper, we make frequent comparisons between two approaches of Koopman learning: snapshot and generator. We restrict our attention to a finite-dimensional lifting function $\phi : \mathcal{X} \to \mathbb{R}^D$, whose components ϕ_i are basis functions belonging to the function class $\mathcal{G}$, together with a dataset $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^n$ consisting of independent and identically distributed samples by ν.

Snapshot-based Approach. Let $\widetilde{\mathbf{F}} : \mathcal{X} \to \mathcal{X}$ be the flow map over a step Δt. Then, starting from $\mathbf{x}_0$, the trajectory is $\mathbf{x}_k = \widetilde{\mathbf{F}}^k(\mathbf{x}_0)$ for $k \in \mathbb{N}_0$. The corresponding Koopman operator family is $(\mathbb{A}^k g)(\mathbf{x}) = g(\widetilde{\mathbf{F}}^k(\mathbf{x})), g \in \mathcal{G},\ k \geq 0$, which induces a trajectory in the observable function space, $g_k = \mathbb{A}^k g, g \triangleq g_0$. That is, instead of evolving the state trajectory $\{\mathbf{x}_k\}$, one may equivalently evolve the observable trajectory $\{g_k\}$. Thus, for vector valued $\phi(\approx \mathbf{g})$, the snapshot-based loss is given by

$$\min_{\mathbf{A}\in\mathbb{R}^{D\times D}} \frac{1}{n}\sum_{i=1}^n \|\phi(\widetilde{\mathbf{F}}(\mathbf{x}_i)) - \mathbf{A}\phi(\mathbf{x}_i)\|_2^2 \ \triangleq\ \frac{1}{n}\sum_{i=1}^n \|s_{\mathbf{A}}(\mathbf{x}_i)\|_2^2. \tag{1}$$

Generator-based Approach. In continuous time, for the dynamical system $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$, the infinitesimal generator $\mathbb{L}$, also referred to as the Lie operator, characterizes the evolution of observables. In particular, generator-level consistency is obtained through $\mathbb{L}g \equiv \nabla_{\mathbf{x}} g \cdot \mathbf{f}$. This follows directly from the chain rule: $\frac{d}{dt}g(\mathbf{x}(t)) = \nabla_{\mathbf{x}} g(\mathbf{x}(t)) \cdot \dot{\mathbf{x}}(t) = \nabla_{\mathbf{x}} g(\mathbf{x}(t)) \cdot \mathbf{f}(\mathbf{x}(t))$. As a result, for a vector-valued observable $\phi(\approx \mathbf{g})$, the generator-based loss is defined by minimizing

$$\min_{\mathbf{L}\in\mathbb{R}^{D\times D}} \frac{1}{n}\sum_{i=1}^n \|\nabla_{\mathbf{x}}\phi(\mathbf{x}_i) \cdot \mathbf{f}(\mathbf{x}_i) - \mathbf{L}\phi(\mathbf{x}_i)\|_2^2 \ \triangleq\ \frac{1}{n}\sum_{i=1}^n \|r_{\mathbf{L}}(\mathbf{x}_i)\|_2^2. \tag{2}$$

Kernel Method for Generator-based Approach. Snapshot-based kernels support one-step rollouts via a dual formulation, whereas generator-based kernels require explicit feature representations. We therefore present the construction of dominant observables for generator-based kernel formulation. Specifically, given training samples $\{\mathbf{x}_i\}_{i=1}^n$ and kernel function, we form the empirical Gram matrix $\mathbf{K} \in \mathbb{R}^{n\times n}$ and, by Mercer's theorem (Mercer, 1909), obtain its eigendecomposition $\mathbf{K} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$. An approximate D-dimensional feature map is then defined as

$$\phi_{\mathrm{KER}}(\mathbf{x}) = \left[\sqrt{\lambda_1}u_1(\mathbf{x}), \ldots, \sqrt{\lambda_D}u_D(\mathbf{x})\right]^\top, \tag{3}$$

where u_j are empirical eigenfunctions associated with the leading eigenvalues λ_j. The Koopman generator $\mathbf{L}_{\mathrm{KER}} \in \mathbb{R}^{D\times D}$ is then estimated by solving $\nabla_{\mathbf{x}}\phi_{\mathrm{KER}}(\mathbf{x})\mathbf{f}(\mathbf{x}) \approx \mathbf{L}_{\mathrm{KER}}\phi_{\mathrm{KER}}(\mathbf{x})$ in the least-squares sense. Finally, a linear decoder $\mathbf{C}$ satisfying $\mathbf{x} \approx \mathbf{C}\phi_{\mathrm{KER}}(\mathbf{x})$ allows rollout as

$$\phi_{\mathrm{KER}}(\mathbf{x}_t) = e^{\mathbf{L}_{\mathrm{KER}}t}\phi_{\mathrm{KER}}(\mathbf{x}_0), \quad \hat{\mathbf{x}}_t = \mathbf{C}\phi_{\mathrm{KER}}(\mathbf{x}_t). \tag{4}$$

Constructing the kernel feature map requires forming and decomposing the Gram matrix $\mathbf{K} \in \mathbb{R}^{n\times n}$, incurring $O(n^3)$ computational cost for each online update. Despite this limitation, kernel-induced features offer highly expressive representations when the kernel matrix is efficiently approximated.

2.3. Problem Setup

We characterize the behavior of Koopman learning when the observable dictionary is fixed. We first formalize the functional setting and approximation structure used in snapshot-based learning. Section 3 derives long-horizon error bounds under this fixed-observable regime, assuming that the dictionary is sufficiently expressive. Section 4 then relaxes this assumption by introducing learnable random Fourier feature observables.

To make these statements precise, we now specify the mathematical setting and assumptions used throughout our analysis. To ensure analytical clarity, we impose regularity conditions on the dynamics and the observable map. Specifically, we assume that the underlying dynamics $\mathbf{f} \in C^2(\mathcal{X};\mathbb{R}^d)$ and observable mapping $\phi \in C^2(\mathcal{X};\mathbb{R}^D) \cap L^2(\nu;\mathbb{R}^D)$ are defined on a compact domain $\mathcal{X}$, with uniformly bounded Jacobian and Hessian, i.e., $M_{\mathcal{X}} \triangleq \sup_{\mathbf{x}\in\mathcal{X}}\{\|J_{\mathbf{f}}(\mathbf{x})\|, \|\nabla^2_{\mathbf{x}}\phi(\mathbf{x})\|\} < \infty$.

Definition 1 (Feature subspace with induced inner product) *Given a lifting function $\phi : \mathcal{X} \to \mathbb{R}^D$, define the feature subspace $\mathcal{V}_\phi \triangleq \{\mathbf{B}\phi \mid \mathbf{B} \in \mathbb{R}^{D\times D}\} \subset L^2(\nu;\mathbb{R}^D)$ with inner product inherited from $L^2(\nu;\mathbb{R}^D)$.*

Throughout Section 3, we assume that ϕ spans a Koopman-invariant subspace, meaning that for all $\mathbf{g} \in \mathcal{V}_\phi$, one has $\mathbb{A}\mathbf{g} \in \mathcal{V}_\phi$. This assumption yields a self-consistent linear approximation of the dynamics within the span of ϕ but does not imply completeness in representing the full nonlinear flow. Indeed, insufficient expressiveness of ϕ may lead to degraded state reconstruction, motivating the following projection operator.

Definition 2 (Orthogonal projection onto $\mathcal{V}_\phi$) *For any $\mathbf{g} \in L^2(\nu;\mathbb{R}^D)$, the orthogonal projection onto $\mathcal{V}_\phi$ is defined as $\Pi_{\mathcal{V}_\phi}\mathbf{g} \triangleq \left(\mathbb{E}_\nu[\mathbf{g}\phi^\top]\right)\Sigma^{-1}\phi$, where $\Sigma \triangleq \mathbb{E}_\nu[\phi\phi^\top]$, $\lambda_{\min}(\Sigma) > 0$.*

When $\mathcal{V}_\phi$ fails to be invariant, the snapshot loss reflects the projection error between the true propagated observable $\mathbb{A}\phi$ and its representation within $\mathcal{V}_\phi$, $(\mathrm{Id} - \Pi_{\mathcal{V}_\phi})\mathbb{A}\phi$. Thus, a finite observables can only represent the partial Koopman action, revealing an intrinsic limitation rooted in the expressiveness of ϕ. Section 4 later addresses this limitation by learning expressive RFF-based observables.

3. Long-Horizon Prediction in Model-Based Environments

Generator-based learning schemes are hypothesized (Azencot et al., 2020; Nayak et al., 2024) to better align with the objective of long-horizon prediction, as the generator loss enforces pointwise consistency with $\mathbf{f}$. In this section, we formalize and theoretically examine this conjecture. We first consider the case of a snapshot-based learning scheme. Let $\mathbf{g}_k = \phi(\mathbf{x}_k)$, and assume that the Koopman operator $\mathbf{A}$ has been estimated through minimization of the snapshot loss $s_{\mathbf{A}}$. Given the true dynamics $\mathbf{x}_{k+1} = \widetilde{\mathbf{F}}(\mathbf{x}_k)$ and the recursive relation $\mathbf{g}_{k+1} = \mathbf{A}\mathbf{g}_k$, the accumulated long-horizon error can be expressed as $\mathbf{g}_K - \mathbf{A}^K\mathbf{g}_0 = \sum_{k=0}^{K-1} \mathbf{A}^{K-1-k} s_{\mathbf{A}}(\mathbf{x}_k)$, $s_{\mathbf{A}}(\mathbf{x}_k) = \mathbf{g}_{k+1} - \mathbf{A}\mathbf{g}_k$. This observation yields the following lemma:

Lemma 3 (Snapshot-based bound) *Under the power-bounded condition $M_K \triangleq \sup_{1\le k\le K} \|\mathbf{A}^k\|_2$, the error accumulates linearly with K. Specifically,*

$$\|\mathbf{g}_K - \mathbf{A}^K\mathbf{g}_0\|_2 \;\le\; M_K\, K \sup_{\mathbf{x}\in\mathcal{X}} \|s_{\mathbf{A}}(\mathbf{x})\|_2.$$

Since $K = T/\Delta t$, the cumulative error bound reflects the effect of the refined temporal discretization. We now consider the generator residual $r_{\mathbf{L}}$ and its corresponding snapshot form

$$s_{\mathbf{L}}(\mathbf{x}) \triangleq \phi(\widetilde{\mathbf{F}}(\mathbf{x})) - e^{\mathbf{L}\Delta t}\phi(\mathbf{x}),$$

which can be viewed as a generator-induced version of the empirical snapshot residual $s_{\mathbf{A}}$. When training is based on $r_{\mathbf{L}}$ and we set $\mathbf{A} = e^{\mathbf{L}\Delta t}$, the K-step prediction naturally satisfies $\mathbf{A}^K = e^{\mathbf{L}T}$. Before deriving the main results, we further establish a key lemma that explicitly connects the generator- and snapshot-based formulations.

Lemma 4 (Generator-based snapshot approximation) *There exists $M_2 < \infty$ (depending only on $M_{\mathcal{X}} = \sup_{\mathbf{x}\in\mathcal{X}}\{\|\nabla^2_{\mathbf{x}}\phi(\mathbf{x})\|, \|J_{\mathbf{f}}(\mathbf{x})\|\}$) such that*

$$s_{\mathbf{L}}(\mathbf{x}) = \Delta t\, r_{\mathbf{L}}(\mathbf{x}) + \mathcal{R}_2(\mathbf{x}, \Delta t), \quad \sup_{\mathbf{x}\in\mathcal{X}} \|\mathcal{R}_2(\mathbf{x}, \Delta t)\|_2 \le M_2\, \Delta t^2. \tag{5}$$

The detailed proof of Lemma 4 is provided in Appendix A.2. This shows that the snapshot discrepancies $s_{\mathbf{L}}$ admit a decomposition involving the generator residuals $r_{\mathbf{L}}$ and some higher-order corrections. With these lemmas, we now establish a multi-step approximation error bound for the generator-induced version as follows:

Theorem 5 (Generator-based bound) *Suppose the system $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ is trained using the generator loss in (2) under the same conditions as Lemma 3. Let the Koopman matrix be constructed as $\mathbf{A} \triangleq e^{\mathbf{L}\Delta t}$. Then, the long-horizon error satisfies*

$$\|\mathbf{g}_K - \mathbf{A}^K\mathbf{g}_0\|_2 \;\lesssim\; M_K\, T \sup_{\mathbf{x}\in\mathcal{X}} \|r_{\mathbf{L}}(\mathbf{x})\|_2 + O(\Delta t).$$

The detailed proof is provided in Appendix A.3. This comparison shows the key distinction: snapshot-based training without generator control yields error growth that depends on the grid resolution ($K = T/\Delta t$). It demonstrates that the accumulated error along trajectories can be controlled by $r_{\mathbf{L}}$, thereby ensuring reliable long-term prediction and stability.

4. Scalable RFF Lifting with Convergence Guarantees

Thus far, our analysis has focused on long-horizon error under the assumption that a sufficiently expressive finite-dimensional observable ϕ is available. However, as noted earlier, the overall performance of Koopman learning is constrained by the expressiveness of the chosen observable space. In particular, it depends on how well the given dictionary ϕ spans a Koopman-relevant subspace $\mathcal{V}_\phi$. A sufficiently expressive dictionary reduces the projection error $(\mathrm{Id} - \Pi_{\mathcal{V}_\phi})\mathbb{A}\phi$, thereby improving long-horizon fidelity.

4.1. Motivation: Structural Challenges of $\mathcal{G}$

The structure of the observable function space $\mathcal{G}$ associated with a given dynamics $\mathbf{f}$ is generally nontrivial. When designing a practically expressive subspace $\widehat{\mathcal{G}}$, we identify two fundamental limitations of common MLP-based and kernel-based methods.

(i) Limited expressiveness due to spectral bias. MLP-based observables ϕ_θ often exhibit spectral bias, favoring low-frequency modes and thereby underrepresenting high-frequency dynamics. This imbalance hinders the construction of an expressive observable capable of representing highly complex dynamical systems.

(ii) Kernel scalability under abundant sampling. Kernel-based approaches such as Kernel Dynamic Mode Decomposition (DMD) effectively capture rich spectral structures but incur cubic complexity in the number of samples, $\mathcal{O}(n^3)$, which limits their scalability in data-abundant regimes.

These structural limitations have motivated the use of random feature lifting in prior studies, such as the RFF-based Extended DMD and kernel generator regression methods (Klus et al., 2020a). However, these approaches typically employed fixed random or kernel feature mappings as static surrogates without explicitly modeling the underlying learning process.

4.2. Proposed Framework: RFF-PIKN

To address these issues, we introduce RFF lifting with a learnable weight matrix $\mathbf{W}_n$. This formulation retains the spectral advantages of MLP-based observables while directly approximating kernel embeddings with linear-time scalability. The detailed construction is as follows:

(i) Random feature approximation. For a shift-invariant kernels, the observable space $\mathcal{G}$ can be interpreted as an RKHS $\widehat{\mathcal{G}}$ induced by $k(\mathbf{x}, \mathbf{y}) = k(\mathbf{x} - \mathbf{y})$. With n samples $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^n$, any $g \in \widehat{\mathcal{G}}$ can be approximated in the finite subspace $\widehat{g}_n(\mathbf{x}) = \sum_{i=1}^n \alpha_i k(\mathbf{x}, \mathbf{x}_i)$. By Bochner's theorem (Rudin, 1962), a shift-invariant kernel admits a RFF approximation $k(\mathbf{x}, \mathbf{y}) \approx \psi(\mathbf{x})^\top \psi(\mathbf{y})$, where $\psi : \mathcal{X} \to \mathbb{R}^{N_\psi}$ and $N_\psi \ll n$. Accordingly, the empirical estimator

$$\widehat{g}_n(\mathbf{x}) = \sum_{i=1}^n \alpha_i k(\mathbf{x}, \mathbf{x}_i) \approx \sum_{i=1}^n \alpha_i \psi(\mathbf{x})^\top \psi(\mathbf{x}_i) = \Big(\sum_{i=1}^n \alpha_i \psi(\mathbf{x}_i) \Big)^\top \psi(\mathbf{x}) = \mathbf{w}_n^\top \psi(\mathbf{x}),$$

where $\mathbf{w}_n \triangleq \sum_{i=1}^n \alpha_i \psi(\mathbf{x}_i)$. By treating $\mathbf{w}_n$ as a trainable parameter, the computational cost shifts from the sample size n to the feature dimension N_ψ. This eliminates the need to invert the kernel matrix. Therefore the method achieves scalability while preserving the expressiveness of the original kernel representation.

(ii) Vector-valued observable parameterization. The formulation naturally extends to vector-valued observables $\mathbf{g} = (g_1, \ldots, g_D)$ by stacking the parameter vectors as

$$\mathbf{W}_n = \begin{bmatrix} \mathbf{w}_{n,1}^\top \\ \vdots \\ \mathbf{w}_{n,D}^\top \end{bmatrix} \in \mathbb{R}^{D \times N_\psi}, \quad \phi(\mathbf{x}) = \mathbf{W}_n \psi(\mathbf{x}).$$

which directly serves as the finite-dimensional observable basis for Koopman approximation.

4.3. SGD Algorithm and Convergence Analysis

RFF-PIKN combines scalability with adaptability for stable Koopman approximation. In a model-based environment that allows i.i.d. resampling, we maintain a moderate sample size n and update $\mathbf{W}_n$ online by SGD with $\mathcal{O}(N_\psi)$ cost. Specifically, we define the following regularized empirical risk (e.g., Koopman-type regularizers such as reconstruction/decoding losses (Lusch et al., 2018; Takeishi et al., 2017; Otto and Rowley, 2019)):

$$\mathcal{L}_n(\mathbf{W}_n^{(t)}, \mathbf{L}_n^{(t)}) \triangleq \frac{1}{n} \sum_{i=1}^{n} \|\nabla_{\mathbf{x}}[\mathbf{W}_n^{(t)} \psi(\mathbf{x}_i^{(t)})] \cdot \mathbf{f}(\mathbf{x}_i^{(t)}) - \mathbf{L}_n^{(t)} \mathbf{W}_n^{(t)} \psi(\mathbf{x}_i^{(t)})\|_2^2 + \lambda \mathcal{L}_{\text{reg}}. \tag{6}$$

To exploit the model-based environment, at each iteration, resample the collocation points $\mathcal{D}^{(t)} = \{\mathbf{x}_i^{(t)}\}_{i=1}^n$, $\mathbf{x}_i^{(t)} \overset{\text{iid}}{\sim} \nu$. Update $(\mathbf{W}_n^{(t)}, \mathbf{L}_n^{(t)})$ by SGD with step size η_t,

$$\begin{aligned} \mathbf{W}_n^{(t+1)} &\leftarrow \mathbf{W}_n^{(t)} - \eta_t \nabla_{\mathbf{W}} \mathcal{L}_n(\mathbf{W}_n^{(t)}, \mathbf{L}_n^{(t)}), \\ \mathbf{L}_n^{(t+1)} &\leftarrow \mathbf{L}_n^{(t)} - \eta_t \nabla_{\mathbf{L}} \mathcal{L}_n(\mathbf{W}_n^{(t)}, \mathbf{L}_n^{(t)}). \end{aligned} \tag{7}$$

This enables efficient adaptation under streaming data. The online update also induces a dynamic alignment between the RFF subspace and the evolving Koopman eigenfunctions. As a result, the framework maintains consistent long-horizon prediction without kernel recomputation or spectral decomposition. We next establish that, under mild smoothness assumptions, the stochastic update of $\mathbf{W}_n^{(t)}$ and $\mathbf{L}_n^{(t)}$ enjoys the same asymptotic convergence properties as standard SGD.

Theorem 6 (Convergence of RFF-PIKN) *Let $\mathbf{f} : \mathcal{X} \to \mathbb{R}^d$ be Lipschitz (as we assume $\mathbf{f} \in C^2(\mathcal{X}; \mathbb{R}^d)$ and $\mathcal{X}$ is compact), and let $\psi : \mathcal{X} \to \mathbb{R}^{N_\psi}$ be fixed. Assume that the step sizes in (7) satisfy $\sum_t \eta_t = \infty$ and $\sum_t \eta_t^2 < \infty$. Then,*

$$(\mathbf{W}_n^{(t)}, \mathbf{L}_n^{(t)}) \to (\mathbf{W}_{\text{RFF}}, \mathbf{L}_{\text{RFF}}) \quad \textit{a.s.},$$

where $(\mathbf{W}_{\text{RFF}}, \mathbf{L}_{\text{RFF}})$ is a stationary point of the true population risk $\mathcal{L}(\mathbf{W}, \mathbf{L}) \triangleq \mathbb{E}_\nu[\mathcal{L}_n(\mathbf{W}, \mathbf{L})]$.

The proof of this result follows from the classical Robbins–Monro stochastic approximation and is provided in Appendix A.4. The learned basis $\{(\mathbf{W}_{\text{RFF}}\psi)_j\}_{j=1}^D$ effectively replaces the empirical eigenfunctions $\{u_j\}_{j=1}^D$ in (3), obtained from the eigendecomposition of the true kernel matrix $(\mathbf{K})_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$. In our setting, the kernel corresponding to the RFF approximation is the RBF kernel $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|_2^2)$. It thus provides a tractable and adaptive approximation of the Koopman learning. We further investigate the significance of this local solution by evaluating its performance relative to the true kernel solution in the subsequent experimental section.

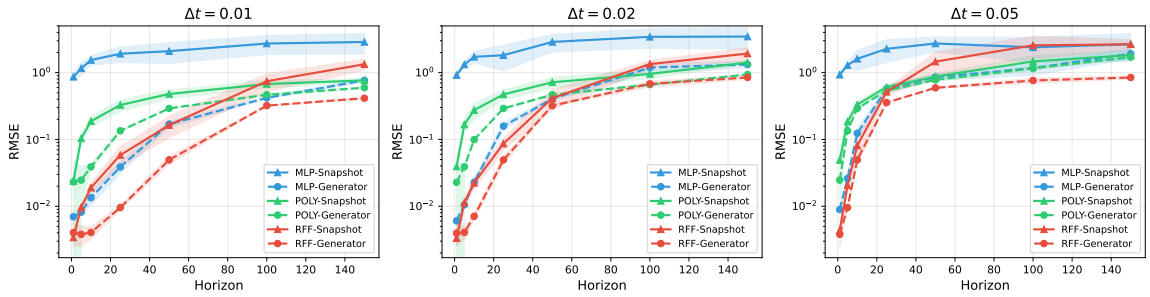

Figure 1: Long-horizon prediction error (RMSE, log-scale) on the Van der Pol (VdP) system. We compare snapshot-based (triangle marker) and generator-based (circle marker) schemes with different feature maps (MLP, polynomial, RFF) under varying step sizes $\Delta t = 0.01, 0.02, 0.05$.

5. Experiments

5.1. Snapshot vs. Generator Formulations for Long-Horizon Prediction

First, we empirically validate Theorem 5 by comparing snapshot- and generator-based methods across three observables—MLP, polynomial, and RFF lifted bases—under varying step sizes (see Figure 1). Generator-based methods consistently yield lower long-horizon errors, demonstrating their advantage. The gain of generator-based approach is most pronounced with expressive MLP encoders and less noticeable for limited-capacity polynomial features. Moreover, as the integration step size Δt increases, this prediction advantage gradually diminishes.

We further evaluate the generality of our Theorem 5 on two representative nonlinear systems: the Duffing oscillator and the double-well attractor. As shown in Table 1, generator-based training consistently outperforms its snapshot-based counterparts across all horizons, and the performance gap widens at longer horizons, reflecting improved long-term stability. Notably, in many cases, the generator formulation even achieves lower errors at $H = 1$, suggesting that enforcing infinitesimal generator consistency enhances both short-term and long-horizon accuracy. These findings reinforce the insight that generator-level learning yields a more coherent representation of the underlying dynamics for Koopman learning.

Table 1: RMSE across horizons and systems (mean $\pm$ std over 5 seeds) with $\Delta t = 0.01$. Best is **bold**, second-best is underlined.

Method	Duffing			Double well		
	H=1	H=50	H=150	H=1	H=50	H=150
MLP–Snapshot	0.73 ± 0.09	1.89 ± 0.49	2.97 ± 0.74	0.85 ± 0.11	2.49 ± 0.55	2.73 ± 1.08
MLP–Generator	$\mathbf{0.00 \pm 0.00}$	0.03 ± 0.00	0.44 ± 0.02	$\mathbf{0.00 \pm 0.00}$	0.05 ± 0.01	0.58 ± 0.08
Poly–Snapshot	$\mathbf{0.00 \pm 0.00}$	0.18 ± 0.02	0.33 ± 0.08	0.01 ± 0.00	0.46 ± 0.07	1.21 ± 0.19
Poly–Generator	0.01 ± 0.01	0.05 ± 0.01	0.35 ± 0.12	0.01 ± 0.00	0.12 ± 0.00	0.61 ± 0.03
RFF–Snapshot	$\mathbf{0.00 \pm 0.00}$	0.46 ± 0.26	1.97 ± 1.11	0.01 ± 0.00	0.31 ± 0.06	2.04 ± 0.85
RFF–PIKN (ours)	$\mathbf{0.00 \pm 0.00}$	$\mathbf{0.02 \pm 0.00}$	$\mathbf{0.21 \pm 0.03}$	$\mathbf{0.00 \pm 0.00}$	$\mathbf{0.03 \pm 0.00}$	$\mathbf{0.33 \pm 0.06}$

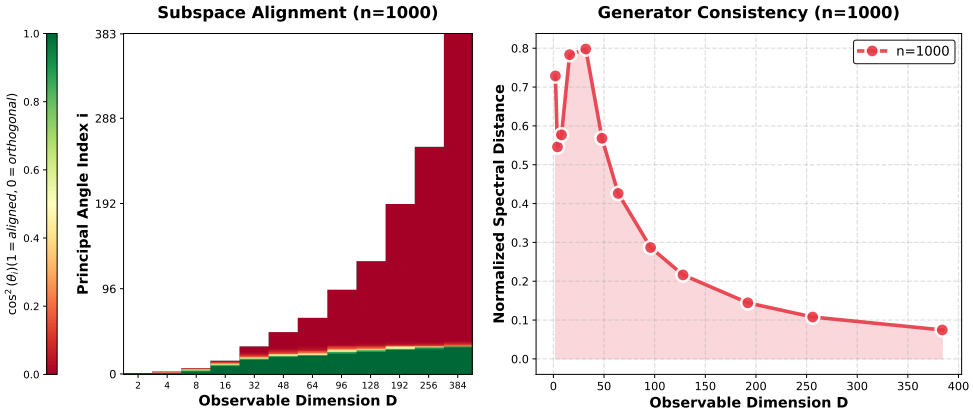

Figure 2: Under the same VdP system, (Left) principal axes obtained through subspace alignment, and (Right) the generator consistency as a function of D, measuring the normalized spectral distance between $\mathbf{L}_{\mathrm{RFF}}$ and $\mathbf{L}_{\mathrm{KER}}$.

5.2. Subspace Alignment and Generator Consistency Test for RFF Approximation

Next, we empirically validate Theorem 6. Namely, we examine the convergence and equivalence between the proposed RFF-PIKN and the conventional kernel–SVD formulation detailed in equations (3) and (4). Fixing $N_\psi = 512$, we vary the sample size n and observable dimension D. For each (n, D), the kernel–SVD baseline is constructed using an RBF kernel $\mathbf{K} \in \mathbb{R}^{n\times n}$ (bandwidth matched to the RFF spectrum). Its rank-D eigendecomposition $\mathbf{K} \approx \mathbf{U}_D \Lambda_D \mathbf{U}_D^\top$ defines features ϕ_{KER}, and the least-squares generator $\mathbf{L}_{\mathrm{KER}}$. Under identical data and bandwidth, the RFF-PIKN model produces $\phi_{\mathrm{RFF}} \triangleq \mathbf{W}_{\mathrm{RFF}}\psi$ and $\mathbf{L}_{\mathrm{RFF}}$, enabling direct comparison via the feature matrices $\Phi_{\mathrm{KER}} \triangleq [\phi_{\mathrm{KER}}(\mathbf{x}_1), \ldots, \phi_{\mathrm{KER}}(\mathbf{x}_n)]^\top$ and $\Phi_{\mathrm{RFF}} \triangleq [\phi_{\mathrm{RFF}}(\mathbf{x}_1), \ldots, \phi_{\mathrm{RFF}}(\mathbf{x}_n)]^\top$.

Evaluation Metrics. We employ two complementary metrics to assess equivalence between the RFF-PIKN and kernel method:

- **Subspace Alignment:** Principal angles θ_i between Φ_{KER} and Φ_{RFF} after canonical correlation whitening, with $\cos\theta_i = \sigma_i$ obtained from the singular values of the whitened cross-covariance.
- **Generator consistency:** Normalized spectral distance $\frac{1}{D}\sum_{i=1}^{D} \frac{|\lambda_i - \mu_i|}{\max(1, |\lambda_i|)}$, where $\{\lambda_i\}$ and $\{\mu_i\}$ are eigenvalues of $\mathbf{L}_{\mathrm{KER}}$ and $\mathbf{L}_{\mathrm{RFF}}$, respectively.

The results in Figure 2 show that the principal axes from subspace alignment and spectral gap from the generator consistency reveal an effective Koopman-invariant subspace of about $D = 30$ dimensions. Beyond this point, increasing the observable dimension D provides little improvement in the subspace alignment, as the remaining axes become nearly orthogonal and dynamically irrelevant. Meanwhile, the normalized spectral distance exhibits a *double-descent* trend, rising initially, peaking near this critical dimension $D = 30$, and then decreasing with larger D.

These behaviors arise because kernel-based eigenfunctions ϕ_{KER} form an orthogonal, energy-ordered basis that captures dominant dynamical modes under the RBF kernel, resulting in a near-decoupled generator $\mathbf{L}_{\mathrm{KER}}$ well-represented within $\sim$30 dimensions. In contrast, random features

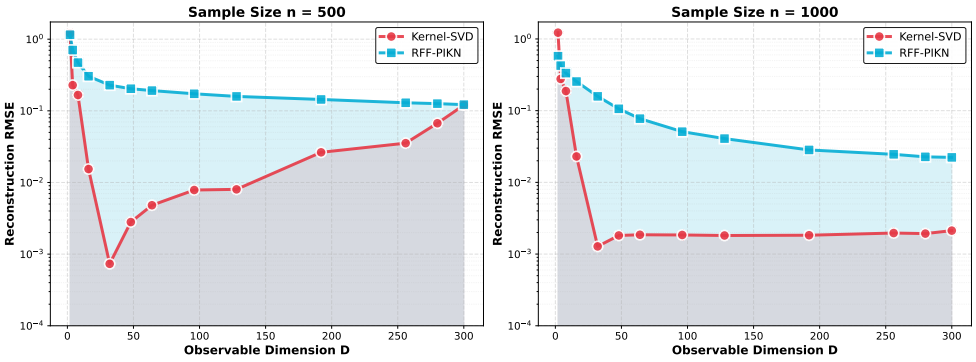

Figure 3: Long-horizon RMSE ($H = 150$) versus feature dimension D, measured under the VdP system. (Left) $n = 500$, (Right) $n = 1000$.

ϕ_{RFF} comprise correlated cosine terms lacking orthogonality, leading to a dense and mode-coupled generator $\mathbf{L}_{\mathrm{RFF}}$. When projected into a low-dimensional space, this coupling amplifies spectral mismatch with $\mathbf{L}_{\mathrm{KER}}$, which gradually diminishes as D increases. These results indicate that subspace alignment does not signify learning saturation but instead reflects a trade-off between the orthogonality and expressiveness of the feature basis and the flexibility of the Koopman generator $\mathbf{L}$.

Insufficiently orthogonal bases require a more adaptive generator; otherwise, trajectories deviate from the Koopman-invariant subspace. This is shown in Figure 3, which presents RMSE versus observable dimension D for both Kernel-SVD and RFF-PIKN under different sample sizes (n=500, 1000). Kernel-SVD shows rapid error reduction up to ∼30 dimensions, beyond which performance saturates; this is consistent with the low-dimensional invariant subspace in Figure 2. In contrast, RFF-PIKN exhibits a smoother decay and higher RMSE floor due to its non-orthogonal, and coupled basis structure. As n and D grow, this gap narrows, suggesting that RFF-PIKN can asymptotically recover the same Koopman subspace given sufficient data and feature capacity.

6. Conclusion

This paper investigated the effectiveness of infinitesimal generator–based Koopman learning for long-horizon prediction through an analytic comparison with snapshot-based losses. To this end, we proposed the Random Fourier Feature–lifted generator loss minimization scheme (RFF-PIKN), which enables scalable generator learning. It addresses the two major challenges of existing observables: the spectral bias of MLPs and the limited scalability of kernel-based method in data-abundant settings. We further examined the alignment between the estimated generator and the underlying kernel basis, revealing clearer learning dynamics. The main bottleneck arises from fitting the generator with a limited basis. Future work includes exploring kernel selection, enhancing random feature expressivity, and extending the framework to stochastic dynamical systems.

Acknowledgments

This work is supported by the Korea Agency for Infrastructure Technology Advancement (KAIA) grant funded by the Ministry of Land, Infrastructure and Transport(Grant RS-2025-02315344).

References

Omri Azencot, Pengzhan Yin, Yuxin He, Michael W. Mahoney, and Andrea L. Bertozzi. Forecasting sequential data using consistent koopman autoencoders. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 465–475, 2020. URL https://www.stat.berkeley.edu/~mmahoney/pubs/icml20-koopman.pdf.

Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Koopman invariant subspaces and finite linear representations of nonlinear dynamical systems for control. *PloS one*, 11(2):e0150171, 2016.

Marko Budišić, Ryan Mohr, and Igor Mezić. Applied koopmanism. *Chaos*, 22(4):047510, 2012. doi: 10.1063/1.4772195. URL https://doi.org/10.1063/1.4772195.

Daniele Castiglione, Danilo Riboni, Matteo Colombini, and Jerome Guzzi. Learning long-horizon predictions for quadrotor dynamics. *arXiv preprint arXiv:2407.12964*, 2024. URL https://arxiv.org/abs/2407.12964.

Nico Gürtler and Georg Martius. Long-horizon planning with predictable skills. *Reinforcement Learning Journal*, 2025. URL https://rlj.cs.umass.edu/2025/papers/RLJ_RLC_2025_136.pdf. Available at rlj.cs.umass.edu.

George E. Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.

Stefan Klus, Feliks Nüske, and Boumediene Hamzi. Kernel-based approximation of the koopman generator and schrödinger operator. *Entropy*, 22(7):722, 2020a. doi: 10.3390/e22070722. URL https://www.mdpi.com/1099-4300/22/7/722.

Stefan Klus, Feliks Nüske, Sebastian Peitz, Jan-Hendrik Niemann, Cecilia Clementi, and Christof Schütte. Data-driven approximation of the koopman generator: Model reduction, system identification, and control. *Physica D: Nonlinear Phenomena*, 406:132416, 2020b. ISSN 0167-2789. doi: https://doi.org/10.1016/j.physd.2020.132416. URL https://www.sciencedirect.com/science/article/pii/S0167278919306086.

Milan Korda and Igor Mezić. Linear predictors for nonlinear dynamical systems: Koopman operator meets model predictive control. *Automatica*, 93:149–160, 2018.

Yuying Liu, Aleksei Sholokhov, Hassan Mansour, and Saleh Nabi. Physics-informed koopman network for time-series prediction of dynamical systems. In *ICLR 2024 Workshop on AI4DifferentialEquations In Science*, 2024.

Bethany Lusch, J Nathan Kutz, and Steven L Brunton. Deep learning for universal linear embeddings of nonlinear dynamics. *Nature Communications*, 9(1):4950, 2018.

James Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 209:415–446, 1909. doi: 10.1098/rsta.1909.0016.

Igor Mezić. Spectral properties of dynamical systems, model reduction and decompositions. *Nonlinear Dynamics*, 41(1):309–325, 2005.

Aniket Nayak, Shahariar Anwar, and Chandra Kambhamettu. Temporally-consistent koopman autoencoders for forecasting dynamical systems. *arXiv preprint arXiv:2403.12335*, 2024. URL https://arxiv.org/abs/2403.12335.

Samuel E Otto and Clarence W Rowley. Linearly recurrent autoencoder networks for learning dynamics. *SIAM Journal on Applied Dynamical Systems*, 18(1):558–593, 2019.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred A. Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5301–5310. PMLR, 2019. URL https://proceedings.mlr.press/v97/rahaman19a.html.

Maziar Raissi and George E. Karniadakis. Physics informed deep learning (part i): Data-driven solutions of nonlinear partial differential equations. *arXiv preprint arXiv:1711.10561*, 2017.

Maziar Raissi, Paris Perdikaris, and George E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

Walter Rudin. *Fourier Analysis on Groups*. Interscience Publishers, 1962.

Tyler Sam, Yudong Chen, and Christina Lee Yu. Overcoming the long horizon barrier for sample-efficient reinforcement learning with latent low-rank structure, 2023. URL https://arxiv.org/abs/2206.03569.

Naoya Takeishi, Yoshinobu Kawahara, and Takehisa Yairi. Learning koopman invariant subspaces for dynamic mode decomposition. In *NeurIPS*, 2017.

Jonathan H Tu, Clarence W Rowley, Dirk M Luchtenburg, Steven L Brunton, and J Nathan Kutz. On dynamic mode decomposition: theory and applications. In *Journal of Computational Dynamics*, volume 1, pages 391–421, 2014.

Matthew O Williams, Ioannis G Kevrekidis, and Clarence W Rowley. A data–driven approximation of the koopman operator: Extending dynamic mode decomposition. *Journal of Nonlinear Science*, 25:1307–1346, 2015.

Chenjun Xiao, Yifan Wu, Chen Ma, Dale Schuurmans, and Martin Müller. Learning to combat compounding-error in model-based reinforcement learning. *arXiv preprint*, 2019. URL https://arxiv.org/abs/1912.11206. arXiv:1912.11206.

Appendix A. Proofs

A.1. Lemma 3

Proof As described in Section 3, let $\mathbf{g}_k = \phi(\mathbf{x}_k)$ denote the lifted state, and the surrogate recursion is given by:

$$\mathbf{g}_{k+1} = \mathbf{A}\mathbf{g}_k,$$

where $\mathbf{A} \in \mathbb{R}^{D\times D}$ denotes the Koopman matrix estimated from data, trained the snapshot-based formulation (1). We evaluate the resulting surrogate dynamics against the true lifted evolution. Specifically, we define the one-step loss induced by the snapshot-based scheme as follows:

$$s_{\mathbf{A}}(\mathbf{x}_k) = \mathbf{g}_{k+1} - \mathbf{A}\mathbf{g}_k.$$

By unrolling the recursion over K steps:

$$\mathbf{g}_K = \mathbf{A}^K\mathbf{g}_0 + \sum_{k=0}^{K-1} \mathbf{A}^{K-1-k} s_{\mathbf{A}}(\mathbf{x}_k),$$

so the total error is:

$$\mathbf{g}_K - \mathbf{A}^K\mathbf{g}_0 = \sum_{k=0}^{K-1} \mathbf{A}^{K-1-k} s_{\mathbf{A}}(\mathbf{x}_k).$$

Taking norm and applying triangle inequality: $\|\mathbf{g}_K - \mathbf{A}^K\mathbf{g}_0\|_2 \le \sum_{k=0}^{K-1} \|\mathbf{A}^{K-1-k}\|_2 \cdot \|s_{\mathbf{A}}(\mathbf{x}_k)\|_2$. Then, under the power-boundedness condition $M_K = \sup_{1\le k\le K} \|\mathbf{A}^k\|_2 < \infty$, we obtain:

$$\|\mathbf{g}_K - \mathbf{A}^K\mathbf{g}_0\|_2 \le M_K \sum_{k=0}^{K-1} \|s_{\mathbf{A}}(\mathbf{x}_k)\|_2 \le M_K K \sup_{\mathbf{x}\in\mathcal{X}} \|s_{\mathbf{A}}(\mathbf{x})\|_2.$$

This concludes the general snapshot-based bound. ■

A.2. Lemma 4

Proof With the estimated Koopman matrix $\mathbf{A} = e^{\mathbf{L}\Delta t}$, the snapshot loss is defined and, for small Δt, can be approximated by

$$s_{\mathbf{L}}(\mathbf{x}) \triangleq \phi(\widetilde{F}(\mathbf{x})) - e^{\mathbf{L}\Delta t}\phi(\mathbf{x}) \approx \phi\left(\mathbf{x} + \Delta t \cdot \mathbf{f}(\mathbf{x})\right) - e^{\mathbf{L}\Delta t}\phi(\mathbf{x}).$$

We perform Taylor expansions of both terms on the right hand side.

(i) Expansion of $\phi(\mathbf{x} + \Delta t \cdot \mathbf{f}(\mathbf{x}))$: By Taylor's theorem with integral remainder, we have

$$\phi(\mathbf{x} + \Delta t \cdot \mathbf{f}(\mathbf{x})) = \phi(\mathbf{x}) + \Delta t \cdot \nabla_{\mathbf{x}}\phi(\mathbf{x})\mathbf{f}(\mathbf{x}) + \mathcal{R}_2^{(1)}(\mathbf{x}, \Delta t),$$

where $\mathcal{R}_2^{(1)}$ depends on the Hessian $\nabla_{\mathbf{x}}^2\phi$ and the Jacobian $J_{\mathbf{f}}$. Since we assumed a uniform bound $M_{\mathcal{X}} = \sup_{\mathbf{x}\in\mathcal{X}}\{\|\nabla_{\mathbf{x}}^2\phi(\mathbf{x})\|, \|J_\phi(\mathbf{x})\|\}$, it follows that $\|\mathcal{R}_2^{(1)}(\mathbf{x}, \Delta t)\|_2 \le C_1\Delta t^2$, for some constant C_1 depending on $M_{\mathcal{X}}$.

(ii) Expansion of $e^{\mathbf{L}\Delta t}\phi(\mathbf{x})$: Using the Taylor expansion of the matrix exponential:

$$e^{\mathbf{L}\Delta t}\phi(\mathbf{x}) = \phi(\mathbf{x}) + \Delta t\mathbf{L}\phi(\mathbf{x}) + \mathcal{R}_2^{(2)}(\mathbf{x}, \Delta t),$$

where the remainder term satisfies $\|\mathcal{R}_2^{(2)}(\mathbf{x}, \Delta t)\|_2 \leq C_2\Delta t^2$, for some constant C_2, by an argument analogous to the expansion of $\phi(\mathbf{x} + \Delta t\,\mathbf{f}(\mathbf{x}))$.

(iii) Subtracting: Now subtract:

$$\begin{aligned} s_{\mathbf{L}}(\mathbf{x}) &\approx \phi(\mathbf{x} + \Delta t \cdot \mathbf{f}(\mathbf{x})) - e^{\mathbf{L}\Delta t}\phi(\mathbf{x}) \\ &= \Delta t\,(\nabla_{\mathbf{x}}\phi(\mathbf{x})\mathbf{f}(\mathbf{x}) - \mathbf{L}\phi(\mathbf{x})) + \left(\mathcal{R}_2^{(1)}(\mathbf{x}, \Delta t) - \mathcal{R}_2^{(2)}(\mathbf{x}, \Delta t)\right) \\ &= \Delta t(\mathbb{L}\phi(\mathbf{x}) - \mathbf{L}\phi(\mathbf{x})) + \left(\mathcal{R}_2^{(1)}(\mathbf{x}, \Delta t) - \mathcal{R}_2^{(2)}(\mathbf{x}, \Delta t)\right). \end{aligned}$$

Now, by the definition of generator residual $r_{\mathbf{L}}(\mathbf{x}) = \frac{d}{dt}\phi(\mathbf{x}) - \mathbf{L}\phi(\mathbf{x}) = \mathbb{L}\phi(\mathbf{x}) - \mathbf{L}\phi(\mathbf{x})$ and defining $\mathcal{R}_2(\mathbf{x}, \Delta t) \triangleq \mathcal{R}_2^{(1)}(\mathbf{x}, \Delta t) - \mathcal{R}_2^{(2)}(\mathbf{x}, \Delta t)$, we get:

$$s_{\mathbf{L}}(\mathbf{x}) = \Delta t \cdot r_{\mathbf{L}}(\mathbf{x}) + \mathcal{R}_2(\mathbf{x}, \Delta t).$$

Since both $\mathcal{R}_2^{(1)}$ and $\mathcal{R}_2^{(2)}$ are uniformly $\mathcal{O}(\Delta t^2)$, this completes the proof. ■

A.3. Theorem 5

Proof Since $\mathbf{A} = e^{\mathbf{L}\Delta t}$, the resulting snapshot residual $s_{\mathbf{L}}(\mathbf{x}_k)$ corresponds to the error induced by the generator-based training. Then, by Lemma 4, we have

$$s_{\mathbf{L}}(\mathbf{x}_k) \leq \Delta t\, r_{\mathbf{L}}(\mathbf{x}_k) + \mathcal{O}(\Delta t^2),$$

and therefore

$$\sup_{\mathbf{x}\in\mathcal{X}} \|s_{\mathbf{L}}(\mathbf{x})\|_2 \leq \Delta t \sup_{\mathbf{x}\in\mathcal{X}} \|r_{\mathbf{L}}(\mathbf{x})\|_2 + \mathcal{O}(\Delta t^2).$$

Substituting into the general bound induced by Lemma 3:

$$\|\mathbf{g}_K - \mathbf{A}^K\mathbf{g}_0\|_2 \leq M_K K \left(\Delta t \sup_{\mathbf{x}} \|r_{\mathbf{L}}(\mathbf{x})\|_2 + \mathcal{O}(\Delta t^2)\right).$$

Recalling that $K = T/\Delta t$, we get:

$$\|\mathbf{g}_K - \mathbf{A}^K\mathbf{g}_0\|_2 \lesssim M_K T \sup_{\mathbf{x}} \|r_{\mathbf{L}}(\mathbf{x})\|_2 + \mathcal{O}(\Delta t).$$

■

A.4. Theorem 6

Proof We consider the parameter pair $(\mathbf{W}_n^{(t)}, \mathbf{L}_n^{(t)})$ and define the regularized population risk as

$$\mathcal{L}(\mathbf{W}_n^{(t)}, \mathbf{L}_n^{(t)}) = \mathbb{E}_{\mathbf{x}\sim\nu}\left[\|\nabla_{\mathbf{x}}[\mathbf{W}_n^{(t)}\psi(\mathbf{x})] \cdot \mathbf{f}(\mathbf{x}) - \mathbf{L}_n^{(t)}\mathbf{W}_n^{(t)}\psi(\mathbf{x})\|_2^2\right] + \lambda\mathcal{L}_{\text{reg}}.$$

At iteration t, we draw a mini-batch $\mathcal{D}^{(t)} = \{\mathbf{x}_i^{(t)}\}_{i=1}^n$ from ν and compute the stochastic gradient $g^{(t)} = \nabla_{(\mathbf{W},\mathbf{L})}\mathcal{L}_n(\mathbf{W}_n^{(t)}, \mathbf{L}_n^{(t)})$, where $\mathcal{L}_n$ denotes the empirical risk in (6), which serves as a finite-sample approximation to the true population risk $\mathcal{L}$. The parameters are updated by standard SGD:

$$(\mathbf{W}_n^{(t+1)}, \mathbf{L}_n^{(t+1)}) = (\mathbf{W}_n^{(t)}, \mathbf{L}_n^{(t)}) - \eta_t g^{(t)}.$$

Since each mini-batch is i.i.d. sampled from ν, the stochastic gradient is an unbiased estimate of the true gradient $\mathbb{E}[g^{(t)}|\mathbf{W}_n^{(t)}, \mathbf{L}_n^{(t)}] = \nabla_{(\mathbf{W},\mathbf{L})}\mathcal{L}(\mathbf{W}_n^{(t)}, \mathbf{L}_n^{(t)})$, and its variance remains bounded due to the Lipschitz continuity of $\mathbf{f}$ and boundedness of ψ. Under these assumptions, together with the Robbins–Monro step-size conditions

$$\sum_t \eta_t = \infty, \qquad \sum_t \eta_t^2 < \infty,$$

the stochastic approximation theorem ensures that the sequence $(\mathbf{W}_n^{(t)}, \mathbf{L}_n^{(t)})$ converges almost surely to a stationary point of the population risk:

$$(\mathbf{W}_n^{(t)}, \mathbf{L}_n^{(t)}) \to (\mathbf{W}_{\text{RFF}}, \mathbf{L}_{\text{RFF}}) \quad \text{a.s.}$$

■