arXiv:2605.18276v1 [stat.ML] 18 May 2026

Geometric Dictionary Learning of Dynamical Systems with Optimal Transport

Thibaut Germain
CMAP, Ecole Polytechnique
thibaut.germain@polytechnique.edu

Sami Chemlal
CMAP, Ecole Polytechnique
samychemlal@yahoo.fr

Rémi Flamary
CMAP, Ecole Polytechnique
remi.flamary@polytechnique.edu

Vladimir R. Kostic
Istituto Italiano di Tecnologia & University of Novi Sad
vladimir.kostic@iit.it

Karim Lounici
CMAP, Ecole Polytechnique
karim.lounici@polytechnique.edu

Abstract

Learning dynamical systems through operator-theoretic representations provides a powerful framework for analyzing complex dynamics, as spectral quantities such as eigenvalues and invariant structures encode characteristic time scales and long-term behavior. However, dynamical operators are typically estimated independently for each system, preventing the discovery of shared structure across related dynamics. To address this limitation, we posit that related dynamical systems lie near a low-dimensional manifold in spectral operator space. Based on this hypothesis, we introduce **DOODL** (*Dynamical OperatOr Dictionary Learning*), a framework that learns a dictionary of characteristic spectral dynamics whose combinations approximate this manifold and yield compact, interpretable embeddings of individual systems. Beyond representation learning, DOODL enables fast and interpretable operator estimation from short and partially observed trajectories by constraining the estimation to the learned operator manifold. Experiments on metastable Langevin dynamics and turbulent plasma simulations demonstrate that DOODL scales to highly complex multiscale regimes while capturing characteristic spectral structure governing the dynamics rather than merely fitting trajectories, achieving errors one to two orders of magnitude lower than independent operator estimation methods in challenging low-data regimes.

1 Introduction

Understanding and modeling dynamical systems (DS) from data is a central problem across scientific and engineering domains. A fundamental challenge is that trajectories of such systems exhibit complex temporal and spatial patterns, arising from multiple interacting time scales and modes. Disentangling these patterns is essential for tasks such as system identification, stability analysis, and control, and is naturally addressed through *spectral decomposition*, which expresses dynamics

Preprint.

in terms of fundamental modes, frequencies, and invariant structures. Koopman operator-theoretic approaches provide a principled framework by describing nonlinear dynamics through linear operators acting on observables [31]. Spectral quantities such as eigenvalues and invariant subspaces encode meaningful information, including time scales and coherent structures.

Estimation of dynamical operators. Learning evolution operators and their spectral structure from data faces several challenges. First, accurate estimation typically requires long and informative trajectories. Second, the operator acts on an *a priori unknown* space of observables that is approximated in practice, for instance with finite non-linear embeddings or kernel methods [63, 29, 30, 38, 33]. As a result, different systems may rely on different observable spaces, and there is no canonical shared structure across systems, making comparison and transfer ill-posed. These challenges are exacerbated in practical settings with short and partial trajectories, often confined to metastable regions. In such regimes, tasks like system identification or long-term prediction become particularly difficult, especially in high-dimensional systems.

Dictionary learning. A natural strategy is to exploit shared structure across systems. Dictionary learning (DL) offers a principled framework for representing data as compositions of reusable atoms, yielding low-dimensional and often interpretable structure [58, 45, 44, 50]. However, classical linear DL methods are designed for static, vector-valued data and do not readily apply to the non-Euclidean space of dynamical operators [26]. A prior work [26] partially addresses this limitation by proposing a DL approach for linear dynamical systems parameterized by symmetric matrices. Yet, this formulation is restricted in scope: it neither captures the broader class of dynamical operators nor accounts for their spectral structure. Related ideas have also been explored in Koopman-based approaches, where dictionaries of observables are learned within Extended Dynamic Mode Decomposition (EDMD) frameworks [42]. However, these methods remain system-specific, operate at the level of observables rather than operators, and do not explicitly learn shared structure across multiple systems.

The geometry of dynamical systems. In parallel, a growing body of work have investigated similarities between dynamical systems through their spectral properties [48, 47, 27, 46, 16, 3, 62, 10]. Notably, [17] introduced a geometry based on optimal transport between spectral components, enabling meaningful notions of distance and interpolation. However, these approaches typically assume fully estimated operators and do not address the problem of learning shared structure from data.

Research gap. These lines of work expose a fundamental gap: existing methods either provide expressive operator-level representations but remain instance-specific, or enable principled comparison without supporting the learning of shared structure. A unified framework that simultaneously captures shared operator structure while enabling transfer and geometric operations remains absent.

Contribution. We introduce DOODL (*Dynamical Operator Dictionary Learning*), a novel framework that operates on a population of related dynamical systems. We posit that related dynamical systems lie near a low-dimensional manifold in spectral operator space. DOODL learns a dictionary of characteristic dynamics (atoms) whose combinations approximate the population manifold and yield shallow vector embeddings of individual systems.

DOODL leverages the spectral decomposition of operators to bind dynamical properties (modes, time scales, frequencies) to a smooth manifold, enabling interpolation and combination of dynamics while preserving dynamical properties. The dictionary is learned by minimizing an optimal transport loss and provides interpretable vector embeddings with physical insight that can be used for downstream tasks such as system identification.

Beyond its embedding capability, DOODL also enables the estimation of dynamical operators from short trajectories by leveraging its learned representation of the population manifold. This behavior is formalized through an oracle inequality (Theorem 1). This estimator opens the door to new applications such as early system identification and detection of regime shifts.

Through experiments on metastable Langevin dynamics and plasma turbulence, we show that DOODL captures characteristic spectral structure governing complex dynamics rather than merely fitting trajectories. The Langevin setting provides a controlled environment to study how learning the spectral manifold of related metastable systems improves the estimation of a new operator from short trajectories. The plasma experiments demonstrate that the learned spectral manifold remains effective in highly turbulent, multiscale regimes characterized by complex spectral interactions, where rapid identification from short trajectory prefixes is critical in fusion-relevant settings.

2 Background

Linear evolution operators & estimation from data. Let $(X_t)_{t\geq 0}$ be a flow on a state space $\mathcal{X}$ with time-homogeneous dynamics. Under mild assumptions, it admits a linear operator representation on an observable space $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$. The transfer operator $A_t : \mathcal{F} \to \mathcal{F}$ evolves an observable f as $[A_t f](x) = \mathbb{E}[f(X_t) \mid X_0 = x]$. The family $(A_t)_{t\geq 0}$ forms a semigroup with infinitesimal generator $G = \lim_{t\to 0^+}(A_t - \mathrm{Id})/t$, so that $A_t = \exp(tG)$ [40, 53].

Assuming $\mathcal{F}$ is a separable Hilbert space and G has a discrete spectrum, it admits the decomposition $G = \sum_j \lambda_j\, g_j \otimes_{\mathcal{F}} f_j$, where (λ_j, f_j, g_j) are eigen-triplets satisfying $Gf_j = \lambda_j f_j$, $G^* g_j = \overline{\lambda_j} g_j$, and $\langle f_j, g_j\rangle_{\mathcal{F}} = \delta_{ij}$. Introducing the spectral projectors $P_j = g_j \otimes_{\mathcal{F}} f_j$, the evolution reads $[A_t f](x) = \sum_j e^{\lambda_j t}\langle f, g_j\rangle_{\mathcal{F}} f_j(x)$, decomposing f into modes $m_j^f = \langle f, g_j\rangle_{\mathcal{F}} f_j$ that evolve as scalar oscillators with timescales $\tau_j = \Re(\lambda_j)^{-1}$ and frequencies $\omega_j = \Im(\lambda_j)/(2\pi)$. This decomposition provides interpretable insights into the system dynamics (e.g., stability, periodicity, mixing), while P_j characterize invariant subspaces and coherent structures. In practice, G is not observed and must be estimated from data. We consider finite-rank estimators $\widehat{G} \in \mathcal{S}_r(\mathcal{H})$ capturing the dominant spectral components. Details on estimation procedures are deferred to Appendix A.1.

Spectral comparison of operators via optimal transport. To compare operators, we propose to use the Spectral-Grassmann OT (SGOT) [17] framework to represent them through their spectral decompositions, capturing both their dynamical time scales and invariant subspaces. Let $G \in \mathcal{S}_r(\mathcal{H})$ admit $G = \sum_{i=1}^{\ell} \lambda_i P_i$, where λ_i are eigenvalues, P_i the associated spectral projectors, and $m_i = \mathrm{rank}(P_i)$ their multiplicities. We associate to G the spectral measure $\mu(G) = \sum_{i=1}^{\ell} \frac{m_i}{m_{\mathrm{tot}}}\, \delta_{(\lambda_i, P_i)}$, with $m_{\mathrm{tot}} = \sum_i m_i$.
Given a divergence $d_{\mathcal{E}}$ on projections and parameters $\eta \in (0,1)$, $q \geq 1$, define the ground cost $C_\eta^q((\lambda, P), (\lambda', P')) = \eta|\lambda - \lambda'|^q + (1-\eta) d_{\mathcal{E}}^q(P, P')$. The spectral optimal transport (SGOT) divergence between G and G' is then

$$d_{\mathcal{S}}(G, G') = \min_{\pi \in \Pi(\mu(G), \mu(G'))} \sum_{i,j} \pi_{ij}\, C_\eta^q\big((\lambda_i, P_i), (\lambda'_j, P'_j)\big), \tag{1}$$

where $\Pi(\mu(G), \mu(G'))$ denotes the set of couplings between spectral measures. This construction, introduced in [17], defines a geometry on operators that jointly captures spectral (eigenvalues) and geometric (invariant subspaces) information, enabling meaningful comparison and interpolation between dynamical systems. We refer to Appendix A.2 for further details and properties.

Dictionary learning. Dictionary learning aims at representing data points $(x_i)_{i\in[N]}$ in a space $\mathcal{X}$ through a small number of atoms $D = (D_\ell)_{\ell\in[K]}$, using a decoding function B such that $x_i \approx \mathrm{B}(\alpha_i; D)$ with $\alpha_i \in \mathbb{R}^K$. The dictionary and codes are typically learned by minimizing a reconstruction objective of the form $\min_{D,(\alpha_i)} \sum_i \mathcal{L}(x_i, \mathrm{B}(\alpha_i; D)) + \mathcal{R}(\alpha_i)$, where $\mathcal{L}$ is a data fitting loss and $\mathcal{R}$ a regularization term. In classical settings, $\mathcal{X}$ is a linear space and B is taken to be linear, $\mathrm{B}(\alpha; D) = \sum_\ell \alpha_\ell D_\ell$. While this paradigm has been extended to structured data (e.g., manifolds or probability measures) by replacing linear combinations with barycentric constructions [25, 22, 60, 61, 54], such formulations have not been developed for dynamical systems represented through operators, whose intrinsic structure is spectral and operator-theoretic rather than Euclidean. See Appendix A.3 for more details.

3 Dynamical Operator Dictionary Learning with Optimal Transport

Problem setting and assumptions. We study dictionary learning of dynamical systems via non-defective operator representations. We assume access to a collection of generators estimated from sampled trajectories, and impose the following conditions: **(A1)** the relevant spectral components of all operators lie in a common finite-dimensional Hilbert space; **(A2)** all eigenvalues are simple (i.e., have multiplicity one). Assumption **(A2)** is standard in practice, since repeated eigenvalues occur with probability zero. Assumption **(A1)** introduces an approximation bias, which can be mitigated by increasing the expressivity of the function space (e.g., through additional Random Fourier Features or richer neural parameterizations).
In the following, we first describe the Riemannian structure of spectral decompositions and the associated optimization tools. We then introduce the Dynamical Operator Dictionary Learning

(DOODL) framework, and finally propose a dictionary-based estimator that improves performance on short trajectories while providing theoretical guarantees.

3.1 A Riemanian manifold of spectral decomposition

We provide a novel characterization of the manifold of spectral decompositions, and derive the tools required to perform gradient-based optimization on this space. While the differential geometry of classical matrix manifolds, including Stiefel and Grassmann manifolds, has been extensively studied and widely applied in optimization and machine learning [1, 7], the manifold structure induced by spectral decompositions has been, to the best of our knowledge, studied only in specific settings. The closest related works focus on real-valued symmetric matrices [2] or bi-orthogonal square matrices [20]; however, these settings differ fundamentally from ours, as we consider low-rank complex matrices, leading to a distinct geometric structure.

A geometric perspective on spectral decomposition. Let $\mathcal{H}$ be a complex p-dimensional Hilbert space and $\mathcal{S}_r(\mathcal{H})$ being the set of non defective operators with rank at most r and simple spectrum. Any $G \in \mathcal{S}_r(\mathcal{H})$ admits a spectral decomposition with matrix representation:

$$\mathbf{G} = \mathbf{R}\,\mathrm{Diag}(\mathbf{\Lambda})\mathbf{L}^* \quad \text{s.t.} \quad \mathbf{L}^*\mathbf{R} = \mathbf{I}_r \;, \tag{2}$$

where $\mathbf{\Lambda} \in \mathbb{C}^r$ are the eigenvalues, $(\mathbf{L}, \mathbf{R}) \in (\mathbb{C}^{p\times r})^2$ are the left/right eigenvectors satisfying the bi-orthogonal condition. However, operators' spectral decomposition are not unique: under the simple spectrum assumption, eigenvectors are defined up to independent rescaling. This induces a natural equivalence relation on spectral decompositions, corresponding to component-wise actions of $(\mathbb{C}^*)^r$ on eigenvectors. We therefore view spectral decompositions as elements of a quotient space

$$\mathcal{M} = \mathcal{N}/(\mathbb{C}^*)^r \quad \text{s.t.} \quad \mathcal{N} = \{(\mathbf{\Lambda}, \mathbf{L}, \mathbf{R}) \mid \mathbf{L}^*\mathbf{R} = \mathbf{I}_r\} \;, \tag{3}$$

where $\mathcal{N}$ denotes the space of eigenvalues and bi-orthogonal eigenvectors. Endowed with this quotient structure, $\mathcal{M}$ admits a smooth manifold structure, and $\mathcal{S}_r(\mathcal{H})$ can be identified with $\mathcal{M}$. All technical details regarding the construction of $\mathcal{N}$ and $\mathcal{M}$ are deferred in Appendix B.

Optimization on the manifold of spectral decomposition. To enable optimization over $\mathcal{M}$, we endow it with a Riemannian metric inherited from the ambient space $\mathcal{N}$. We consider a metric that is invariant under the action of $(\mathbb{C}^*)^r$, ensuring that it is well defined on the quotient manifold $\mathcal{M}$. This construction yields consistent notions of tangent spaces, gradients, and vector transports. In practice, optimization proceeds via Riemannian gradient methods: Euclidean gradients are projected onto the tangent space, and updates are mapped back to the manifold through a suitable retraction. From a computational standpoint, given the Euclidean gradient, a Riemannian gradient step has complexity $\mathcal{O}(r^3 + r^2n^2)$ corresponding to matrix inversions and multiplications. In the typical low-rank regime ($r \ll n$), the inversion cost is negligible, and the overall complexity is effectively governed by the matrix multiplications, making the approach suited for large-scale optimization. We defer the explicit expressions of the metric, tangent space, gradient projection, and retraction to Appendix B.

3.2 Dynamical Operator Dictionary Learning (DOODL)

For conciseness of notation, we denote by $\mathbf{G}$ the matrix representation of an operator $G \in \mathcal{S}_r(\mathcal{H})$, and by $\boldsymbol{\mathcal{G}} = (\mathbf{\Lambda}, \mathbf{L}, \mathbf{R}) \in \mathcal{N}$ its spectral decomposition

Problem formulation & data fitting loss. Given N operators with spectral decompositions $\{\boldsymbol{\mathcal{G}}_i\}_{i\in[N]} \subset \mathcal{N}$, we aim to learn a dictionary of d atoms $\overline{\boldsymbol{\mathcal{G}}} \in \mathcal{N}^d$ by solving

$$\min_{\{\boldsymbol{\alpha}_i\}_{i\in[N]} \subset \mathbf{\Delta}^{d-1},\, \overline{\boldsymbol{\mathcal{G}}} \in \mathcal{N}^d} \quad \frac{1}{N}\sum\nolimits_{i\in[N]} d_{\mathcal{S}}(\mathrm{B}(\boldsymbol{\alpha}_i, \overline{\boldsymbol{\mathcal{G}}}), \boldsymbol{\mathcal{G}}_i) \;, \tag{4}$$

where $\boldsymbol{\alpha}_i \in \mathbf{\Delta}^{d-1}$ are the coefficients associated to the operator atom $\overline{\boldsymbol{\mathcal{G}}_j}$ in the dictionary $\overline{\boldsymbol{\mathcal{G}}}$ and $\mathbf{\Delta}^{d-1}$ the $(d-1)$-simplex, and B is the model reconstructing the operator from its coefficients and the dictionary. The data fitting loss $d_{\mathcal{S}}$ corresponds to the optimal transport divergence defined in eq. (1), with a cost function parametrized with $q = 2$ and a spectral projector divergence $d_{\mathcal{E}}$ that can be chosen depending on the learning task. The dictionary learning problem 4 aims to minimize reconstruction error of operators parametrized and compared through their spectral decompositions. Unlike classical dictionary learning which includes a sparsity regularization term, we constrain coefficients to lie in the simplex, yielding convex combinations of atoms.

Reconstruction model. The goal of the reconstruction model is to recover a spectral decomposition from coefficients $\boldsymbol{\alpha}$ given a dictionary $\overline{\boldsymbol{\mathcal{G}}}$. Classical DL approaches rely on a linear reconstruction, i.e., a weighted sum of dictionary atoms [45, 44]. While natural and effective for linear operators, such models fail to capture the nature of evolution operator, as shown by our numerical experiments. In contrast, the SGOT divergence eq. (1) naturally compares evolution operators through their spectral decompositions [17] suggesting a nonlinear DL strategy at the level of distributions, where reconstruction is performed via Wasserstein barycenters. Such approaches, based on free-support optimal transport barycenters [14], have been explored in [54]. However, despite their geometric fidelity, these methods have substantial computational overhead limiting their practical applicability. A discussion on Wassersterin barycenter with SGOT can be found in Appendix C.1. In a similar fashion, DL methods on Riemannian manifolds [12, 41, 23], typically consider approximation of geodesic barycenters as reconstruction models to avoid the computational burden of Riemannian geometry [25].

Our approach combine the strength of existing methods; we propose a nonlinear model that performs a linear combination of spectral components followed by a projection onto the manifold of spectral decompositions $\mathcal{N}$. This approach leverage efficiency of linear models and accounts for the underlying manifold. Formally, given a dictionary $\overline{\boldsymbol{\mathcal{G}}} \in \mathcal{N}^d$ and coefficients $\boldsymbol{\alpha} \in \boldsymbol{\Delta}^{(d-1)}$, the reconstruction is:

$$\mathrm{B}^{\mathrm{p}}(\boldsymbol{\alpha}; \overline{\boldsymbol{\mathcal{G}}}) \triangleq \mathrm{P}_{\mathcal{N}}(\textstyle\sum_{j\in[d]} \alpha_j \overline{\boldsymbol{\Lambda}}_j, \sum_{j\in[d]} \alpha_j \overline{\mathbf{L}}_j, \sum_{j\in[d]} \alpha_j \overline{\mathbf{R}}_j) \; . \tag{5}$$

The projection operator, $\mathrm{P}_{\mathcal{N}}(\boldsymbol{\Lambda}, \mathbf{L}, \mathbf{R}) \triangleq (\boldsymbol{\Lambda}, \arg\min_{\tilde{\mathbf{L}} \text{ s.t. } \tilde{\mathbf{L}}^*\mathbf{R}=\mathbf{I}_r} \|\tilde{\mathbf{L}} - \mathbf{L}\|_F^2, \mathbf{R})$, restores feasibility by projecting the left eigenvectors to satisfy the bi-orthogonality constraint. The construction is asymmetric, as it leaves eigenvalues and right eigenvectors unchanged. For dynamical systems, this asymmetry is often well-justified since right eigenvectors encode the dominant dynamic modes. In this sense, the model can be viewed as a smooth local retraction on $\mathcal{N}$ around $(\boldsymbol{\Lambda}, \mathbf{R})$, see Appendix B. Compared to Wasserstein barycenter-based reconstruction, the proposed model is not invariant to permutations of spectral components due to the linear aggregation step. In practice, this sensitivity can be mitigated by increasing the dictionary size, which reduces ordering effects. On the other hand, its main advantage is its computational efficiency: it avoids any inner optimization or optimal transport problem, while still incorporating manifold structure through a single projection step.

Optimization scheme. We adopt a stochastic inexact block-coordinate strategy similar to the one proposed in [45] for euclidean spaces. Given a batch of size b, we perform two successive steps:

1. *Coefficient estimation:* fix the dictionary and optimize $\{\boldsymbol{\alpha}_i\}_{i\in[b]}$ via gradient descent (with softmax parametrization). Since the problem is not convex, we use an initialization of based on proximity to dictionary atoms, i.e. $\boldsymbol{\alpha}_i \propto \{-d_{\mathcal{S}}(\boldsymbol{\mathcal{G}}_i, \overline{\boldsymbol{\mathcal{G}}}_j)\}_{j\in[d]}$ to start the optimization in a relevant region of the parameter space.
2. *Dictionary update:* fix the coordinates $\{\boldsymbol{\alpha}_i\}_{i\in[b]}$ and update the dictionary with a Riemannian gradient step on $\mathcal{N}^d$ from the objective on the batch. We leverage the envelope theorem to ignore implicit gradients through $\{\boldsymbol{\alpha}_i\}_{i\in[b]}$.

Further details on the optimization scheme can be found in Appendix C. Since operators are typically low-rank with parametrization on simplex, the main computational cost arises in the *coefficient estimation*, due to repeated evaluations of the reconstruction model.

The DOODL framework integrates the strengths of existing DL methods. Its reconstruction model combines efficiency with geometric fidelity which enables the learning of operator dictionaries through Riemannian optimization, where gradients are guided by an optimal transport divergence that captures the geometry of spectral decompositions central to evolution operators.

3.3 Dictionary-based Operator Estimation from Short Trajectories

DOODL can be used to estimate operators from short trajectories by restricting the search space to the image of the reconstruction model B^{p} induced by a learned dictionary. This low-dimensional constraint regularizes estimation in low-data regimes where classical estimators are unreliable. Given a trajectory $(X_t)_{t\in[n]}$, we estimate the operator by solving

$$\min \boldsymbol{\alpha} \quad \tfrac{1}{n} \textstyle\sum_{t\in[n]} \|z_{t+1} - \mathrm{Op}[\mathrm{B}^{\mathrm{p}}(\boldsymbol{\alpha}; \overline{\boldsymbol{\mathcal{G}}})]^* z_t\|_{\mathcal{H}}^2 \; , \tag{6}$$

where $z_t = \phi(X_t)$ denotes the feature representation of X_t in the Hilbert space $\mathcal{H}$, and Op maps a spectral decomposition to its associated operator. This corresponds to empirical risk minimization

constrained to the learned operator manifold. We provide statistical guarantees below showing improved estimation in short-trajectory regimes.

Statistical guarantees. Let $\bar{\boldsymbol{\mathcal{G}}}^*$ denote the optimal dictionary solving (4), and $\hat{\boldsymbol{\mathcal{G}}}$ its empirical estimate. Define $G_{\boldsymbol{\alpha}} := \mathrm{Op}[\mathrm{B}(\boldsymbol{\alpha}; \hat{\boldsymbol{\mathcal{G}}})]$ and $\mathcal{G}(\hat{\boldsymbol{\mathcal{G}}}) := \{G_{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \Delta^{d-1}\}$. Given a stationary test trajectory $(X_t)_{t=1}^{n+1}$ independent of the training data, we define the empirical and population risks

$$\ell_t(G) := \|\mathrm{z}_{t+1} - G^* \mathrm{z}_t\|_{\mathcal{H}}^2, \quad \widehat{\mathrm{R}}(\boldsymbol{\alpha} \mid \hat{\boldsymbol{\mathcal{G}}}) := \tfrac{1}{n} \textstyle\sum_{t=1}^{n} \ell_t(G_{\boldsymbol{\alpha}}), \quad \mathrm{R}(\boldsymbol{\alpha} \mid \hat{\boldsymbol{\mathcal{G}}}) := \mathbb{E}_\pi[\ell_t(G_{\boldsymbol{\alpha}})], \tag{7}$$

with minimizer $\hat{\boldsymbol{\alpha}} \in \arg\min_{\boldsymbol{\alpha}} \widehat{\mathrm{R}}(\boldsymbol{\alpha} \mid \hat{\boldsymbol{\mathcal{G}}})$. We set $\kappa := \sup_x \|\phi(x)\|_{\mathcal{H}}$, $M := \sup_j \|\mathrm{Op}[\hat{\boldsymbol{\mathcal{G}}}_j]\|_{\mathrm{op}}$, $m \le d-1$ the intrinsic dimension of $\mathcal{G}(\hat{\boldsymbol{\mathcal{G}}})$ under $\|\cdot\|_{\mathrm{op}}$ (diameter $\le 2M$, since $\|G_{\boldsymbol{\alpha}}\|_{\mathrm{op}} \le M$ for all $\boldsymbol{\alpha}$), $\sigma^2_{\hat{\boldsymbol{\mathcal{G}}}} := \sup_{G \in \mathcal{G}(\hat{\boldsymbol{\mathcal{G}}})} \mathrm{Var}_\pi(\ell_t(G))$ and $\bar{\sigma}^2_{\mathcal{G}} := \sup_{G \in \mathcal{G}(\hat{\boldsymbol{\mathcal{G}}})} \mathbb{E}_\pi[\ell_t(G)^2]$ the uniform loss variance and second moment. For β-mixing trajectories with block length τ, let $n_{\mathrm{eff}} := \lfloor n/(2\tau) \rfloor$ and $\bar{\sigma}^2_{\mathrm{eff}} := \sigma^2_{\hat{\boldsymbol{\mathcal{G}}}} + 4\bar{\sigma}^2_{\mathcal{G}} \sum_{s=1}^{\tau-1} \beta(s)$. We need the following assumption.

(WC) *Well-conditioned dictionary:* The right eigenvectors $\{\bar{\mathbf{R}}_j\}_{j \in [d]} \subset \mathbb{C}^{p \times r}$ are in general position, i.e. no convex combination $\sum_{j \in [d]} \alpha_j \bar{\mathbf{R}}_j$ is rank-deficient for $\boldsymbol{\alpha} \in \Delta^{d-1}$.

This assumption is analogous to task diversity conditions in meta-learning [59], ensuring that the dominant dynamical modes span a sufficiently rich subspace. It implies the uniform lower bound $\sigma_{\min}(\sum_{j \in [d]} \alpha_j \bar{\mathbf{R}}_j) \ge c > 0$ for all $\boldsymbol{\alpha} \in \Delta^{d-1}$. We denote by L_{dec} the Lipschitz constant of the decoder; bounds for B^{p} is given in Lemma 6.

Theorem 1 (Oracle inequality)**.** *Let* **(WC)** *be satisfied. Suppose the test trajectory $(X_t)_{t \in [n+1]}$ is stationary, β-mixing, and independent of the training data. Then for any $\delta \in (0,1)$, with probability at least $1 - \delta - 2(n_{\mathrm{eff}} - 1)\beta(\tau)$,*

$$\mathrm{R}(\hat{\boldsymbol{\alpha}} \mid \hat{\boldsymbol{\mathcal{G}}}) - \inf_{\boldsymbol{\alpha}} \mathrm{R}(\boldsymbol{\alpha} \mid \bar{\boldsymbol{\mathcal{G}}}) \le \mathrm{bias}(\hat{\boldsymbol{\mathcal{G}}}) + c' \sqrt{\frac{\bar{\sigma}^2_{\mathrm{eff}} \cdot m \log\big(M L_{\mathrm{dec}}(1+M)\kappa^2 \sqrt{n_{\mathrm{eff}}}\big) + \log(1/\delta)}{n_{\mathrm{eff}}}} \tag{8}$$

where $c' > 0$ is a numerical constant and $\mathrm{bias}(\hat{\boldsymbol{\mathcal{G}}}) := \inf_{\boldsymbol{\alpha}} \mathrm{R}(\boldsymbol{\alpha} \mid \hat{\boldsymbol{\mathcal{G}}}) - \inf_{\boldsymbol{\alpha}} \mathrm{R}(\boldsymbol{\alpha} \mid \bar{\boldsymbol{\mathcal{G}}}^*) \ge 0$.

Discussion. The bound separates two terms. The *coordinate fitting* term controls estimation of the low-dimensional embedding coefficients $\hat{\boldsymbol{\alpha}}$ from short trajectories, with complexity governed by the intrinsic dimension m of the learned operator manifold rather than the ambient dictionary size d. The learned dictionary therefore acts as a transferable inductive bias restricting estimation to dynamically plausible operators, which is particularly advantageous in short-trajectory regimes where unconstrained operator estimation is statistically unstable. The effective sample size n_{eff} reflects the mixing properties of the dynamics, while the variance term $\sigma^2_{\hat{\boldsymbol{\mathcal{G}}}}$ measures how well the learned manifold captures the observed dynamics. Finally, the dictionary bias quantifies the approximation gap induced by the learned manifold and vanishes asymptotically as $n_{\mathrm{train}} \to \infty$ (Proposition 7). Full proofs are deferred to Appendix F.

4 Numerical Experiments

Our experiments are twofold: (i) metastable Langevin dynamics, where we study operator estimation from short non-equilibrium trajectories and detection of regime switches; and (ii) turbulent plasma dynamics relevant to nuclear fusion, where we investigate whether DOODL can turn high-dimensional turbulent simulations into low-dimensional spectral coordinates supporting operator recovery and early regime identification.

4.1 Dictionary Learning on Langevin dynamics

Langevin dynamics in science. Langevin dynamics provides a fundamental framework for modeling stochastic systems driven by both deterministic forces and environmental noise. It plays a central role in scientific fields such as molecular dynamics and statistical physics, where key quantities are inherently dynamical and spectral: metastable states, transition pathways, relaxation timescales, and invariant distributions [55, 56]. In practice, reliably estimating the associated evolution operators is challenging because trajectories are often short, high-dimensional, partially observed, and far from

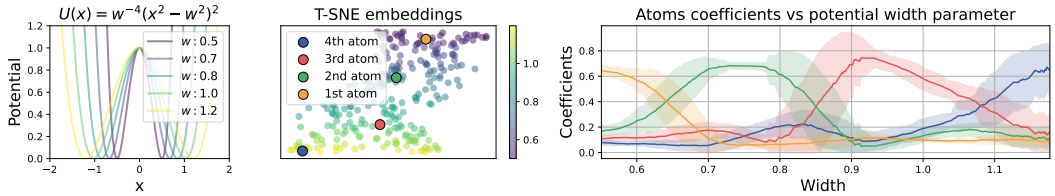

Figure 1: (left) Two-wells Langevin potential for different widths. (middle) T-SNE with SGOT of dictionary atoms and training operators colored by width. (right) Average operators' reconstruction coefficients depending on potential width.

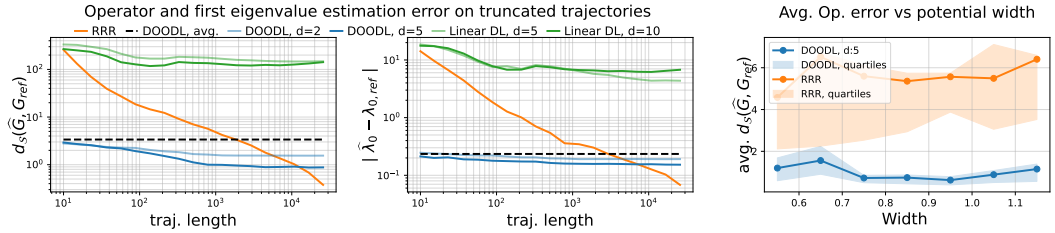

Figure 2: Comparison of operator estimator on truncated trajectories, including individual RRR, linear DL and DOODL. Error to ground truth in SGOT divergence between operators (left) and absolute error between first eigenvalues (middle). (right) Average SGOT error per potential width.

equilibrium [6, 64]. In this experiment, we investigate whether the estimated manifold of dynamical systems through geometric dictionary learning can overcome these limitations.

Data simulation & operator estimation. We consider a one-dimensional damped Langevin dynamics governed by $\mathrm{d}X_t = \nabla U_w(X_t) + \sqrt{2\sigma}\mathrm{d}B_t$ with $\{B_t\}_{t>0}$ a Brownian motion, $\sigma > 0$ its temperature, and a two well potential $U_w(x) \triangleq w^{-4}(x^2 - w^2)^2$ where $w > 0$ controls the distance between the two wells and their respective width, see Fig.1(left). We uniformly sampled 256/256 train/test potential parameters $w \in [0.5, 1.2]$ and generates trajectories of 40k samples at 100Hz according to their Langevin dynamics. The operators are estimated via Reduced Rank Regression (RRR) [34] with a fixed rank fixed of three, and an observable space approximating the RBF kernel via 400 Random Fourier Features [4] on sliding windows of 50 samples. Settings are detailed Appendix D.

Learning the geometry of Langevin dynamics. On the training set, we learn DOODL dictionaries eq. (4) with 2 to 5 atoms, using the SGOT divergence with $\eta = 0.25$ and the log-Martin metric for the spectral projector term. For four atoms, Figure 1 (middle) shows a 2D t-SNE embedding of SGOT distances of training operators and learned atoms, with samples colored by the potential width. Figure 1 (right) reports the corresponding activation coefficients. These results reveal that this family of one-parameter Langevin dynamics lies on a low-dimensional (approximately 1D) manifold in operator space. The learned DOODL atoms align along this structure, yielding smooth and ordered activations as the width parameter varies. This shows that DOODL recovers a coherent low-dimensional organization of the underlying dynamics, where nearby systems share smoothly varying operator representations.

Dictionary-based operator estimation from short trajectories. With previously trained dictionaries, we estimate operators from test trajectories truncated between 10 and 40k samples (20 log-spaced steps), with DOODL in eq. (6). We compare against individual RRR estimators, linear DL baselines, and the average reconstruction model on the training set eq. (5). Figure 2-(left,middle) reports reconstruction error with SGOT divergence and the error on the leading eigenvalue respectively. DOODL outperforms all baselines up to 10k samples, achieving a SGOT error up to two orders of magnitude lower in the short-trajectory regime. This shows that constraining estimation to the learned spectral manifold preserves physically meaningful dynamical structure rather than merely fitting trajectories. Note that for long enough trajectories direct operator estimation RRR performs better because of the bias term in Eq. (8). Linear DL performs worst, suggesting that naive dictionary learning approaches that ignore the intrinsic geometry of operator space fail to recover dynamical properties. This trend is further confirmed in Figure 2-(right), where trajectory length is fixed to 1.2k samples: DOODL yields lower variance across all potential widths compared to RRR, indicating that the learned operator manifold stabilizes operator estimation. Further results are in Appendix D.

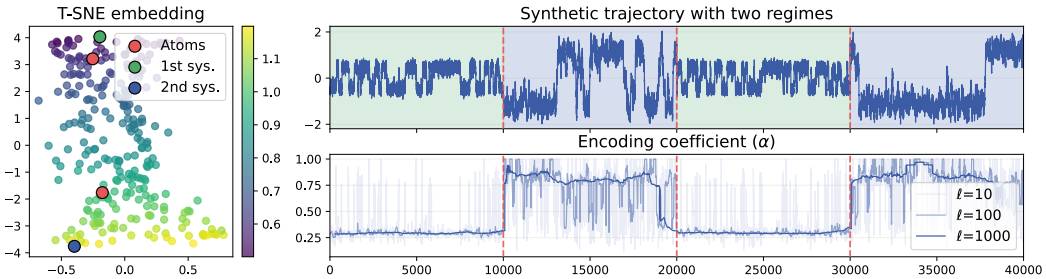

Figure 3: Regime-switch detection from operator embeddings. (left) SGOT geometry of training operators and learned atoms. (right-top) Trajectory with three switches. (right-bottom) Rolling DOODL coefficient for different window lengths.

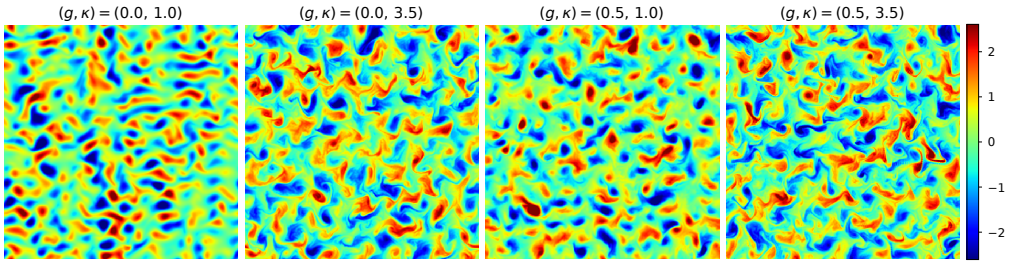

Figure 4: Density fluctuations across plasma regimes for varying (g, κ).

System identification through dictionary-based operator embeddings. Estimating operators from short trajectories naturally enables the detection of regime changes. We illustrate this on a system alternating between two Langevin regimes with wide and narrow potential widths. Figure 3-(top right) shows a representative trajectory with switching events. Considering a DOODL dictionary with two atoms (Figure 3, left), we track the first embedding coefficient of operators estimated over sliding windows of length $\ell \in 10, 100, 1000$. The resulting coefficient trajectories (Figure 3, bottom right) exhibit sharp discontinuities at switching times, with reduced variance as ℓ increases. This demonstrates that the manifold approximation induces coordinate system capable of accurately tracking abrupt changes in the underlying dynamics directly from short trajectory windows.

4.2 Dictionary Learning on plasma dynamics

Plasma dynamics and simulation. We consider a reduced two-field model for edge plasma turbulence with density and electrostatic potential. The parameter, g, controls the interchange coupling between density and vorticity while κ controls the imposed density-gradient drive [18, 19]. Figure 4 shows examples of density fluctuations on different values of g and κ. Increasing κ changes the fluctuation scale and contrast, while increasing g makes the structures more distorted. See Appendix E.1 for details. Estimating such control parameters κ and g from plasma data is a challenge in practice. Turbulence, instabilities, and anomalous transport make quantitative modeling difficult [5]. Even in reduced 2D models, transport parameters require statistical calibration and can vary between regimes [13]. This motivates data-driven reduced-order representations that extract coherent dynamical structures from high-dimensional plasma trajectories [15, 57].
The goal of this experiment is to test whether DOODL can map high-dimensional turbulent Tokam2D [21] simulations into a low-dimensional spectral coordinate system that enables rapid system identification from short turbulent trajectories.
We generate plasma dynamics with Tokam2D [21], a reduced 2D fluid solver that evolves density and vorticity fields in a plane transverse to the magnetic field [18, 19]. In the gradient-driven configuration, we uniformly sample $M = 400$ dynamics with parameters (g, κ) in $[0, 0.5] \times [1, 3.5]$ keeping other numerical parameters fixed. Each system is simulated on a 128×128 grid for $T = 4093$ time steps with $\Delta t = 0.05$. We did $80/20$ train/test split over simulations.

Operator estimations from time series. The learning plasma dynamic operators is in two steps: (i) learning a common latent space for individual operator estimation, (ii) estimating the operator through

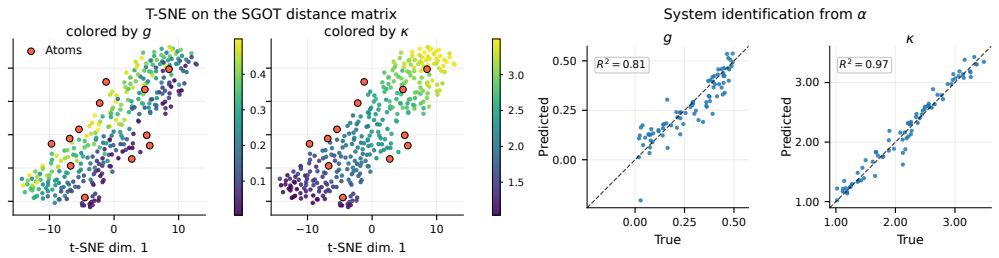

Figure 5: Geometry of turbulent plasma dynamics. Left: SGOT geometry organizes plasma regimes by (g, κ). Right: regression of physical parameters from barycentric embeddings α_m.

their resolvent. The common latent space is learned from sampled trajectories that are first embedded with POSEIDON-B, a foundational model for PDE [24], then fed to two MLP heads applied to consecutive states. The first head produces a 64-dimensional latent state, while the second is only used during training. The heads are trained with a contrastive loss inspired by [36]. Since the systems have different turbulence levels and loss scales, we train the heads with a GradNorm-weighted multitask strategy [11]. From the resulting latent trajectories, we estimate a individual rank-40 operators with a generator-resolvent (GR) estimator [35]. Details are given in Appendix E.

Operator dictionary and recovered manifold. We apply DOODL on the learned operators with $d = 12$ atoms and the SGOT divergence with log-Martin and $\eta = 0.9$. Each system is represented by embeddings $\alpha_m \in \Delta^{d-1}$ which are used for regression on the parameters. T-SNE visualization in Figure 5 shows that the SGOT geometry recovers, only from data, the structure of the 2D manifold of the plasma regimes corresponding to changing both g and κ. The learned atoms lie on the envelope of this operator manifold, consistent with their role as dictionary prototypes. Regressing from the α_m gives $R^2 = 0.81$ for g and $R^2 = 0.97$ for κ on held-out regimes, showing that the DOODL embedding encode the physical parameters.

Early operator recovery and system identification. We test whether the learned dictionary helps recover the plasma operator from short trajectories. For each test trajectory length, we estimate the operator with DOODL and compare it to the reference operator estimated from the full trajectory. Figure 6 (top) shows that the constrained estimate has better performances compared to direct GR estimation for length up to 3k time steps. This demonstrates that constraining estimation to the learned spectral manifold enables reliable identification of complex plasma regimes long before direct operator estimation becomes stable. Once the spectral manifold is learned, inference reduces to estimating a small number of barycentric coordinates rather than recovering an unconstrained high-dimensional operator from scratch, yielding a lightweight reduced-order representation of turbulent plasma dynamics.

We now test whether DOODL coordinates estimated from short trajectories enables system identification. Figure 6 (bottom) reports the R^2 obtained when predicting g and κ from coordinates estimated at each length. The two parameters behave differently. κ is recovered from short trajectories, suggesting that its effect is encoded in leading spectral modes that emerge rapidly. In contrast, g improves more gradually and requires longer trajectories, consistent with results in Appendix E.6 where it is associated with slower and higher-order spectral components. This indicates that the learned operator geometry captures how physical parameters are distributed across spectral time scales in turbulent plasma dynamics. Such lightweight operator representations are particularly attractive in fusion-relevant turbulent plasma regimes where rapid identification from short trajectory prefixes is critical. More broadly, our results support the view that turbulent physical systems admit a low-dimensional manifold representation.

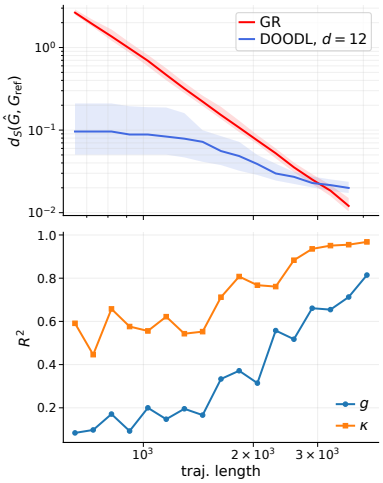

Figure 6: Early operator recovery and parameter identification from short plasma trajectories. Top: operator estimation. Bottom: parameter regression from DOODL embeddings.

5 Conclusion

We introduced DOODL, a geometric dictionary learning framework for dynamical systems based on the hypothesis that related systems lie near a low-dimensional spectral manifold in operator space. DOODL combines spectral theory with optimal transport and Riemannian dictionary learning to construct compact operator representations preserving key dynamical properties. DOODL constrains estimation to the learned spectral manifold, enabling fast, interpretable, and data-efficient reasoning about complex dynamics.

Despite these promising results, several limitations remain. In particular, as is well known in dictionary learning, the underlying optimization problem is non-convex. In our setting, this challenge is further exacerbated by the manifold constraints, motivating the coefficient initialization strategy that we propose. As well, the learned manifold is static and does not adapt online to evolving systems or distribution shifts. Future work will investigate continual and adaptive variants of DOODL, as well as tighter integration between operator estimation and geometric representation learning.

6 Acknowledgments

This project received funding from the European Union's Horizon Europe research and innovation program under grant agreement 101120237 (ELIAS), Fondation de l'Ecole Polytechnique, Hi! PARIS, the French National Research Agency (ANR) through France 2030 program (ANR-23-IACL-0005 and ANR-25-PEIA-0005), NextGenerationEU and MUR PNRR project PE0000013 CUP J53C22003010006 "Future Artificial Intelligence Research (FAIR)".

References

[1] P-A Absil. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2008.

[2] P-A Absil and Jérôme Malick. Projection-like retractions on matrix manifolds. *SIAM Journal on Optimization*, 22(1):135–158, 2012.

[3] Bijan Afsari and René Vidal. Distances on spaces of high-dimensional linear stochastic processes: a survey. In *Geometric Theory of Information*, pages 219–242. Springer, 2014.

[4] Haim Avron, Michael Kapralov, Cameron Musco, Christopher Musco, Ameya Velingker, and Amir Zandieh. Random fourier features for kernel ridge regression: Approximation bounds and statistical guarantees. In *International conference on machine learning*, pages 253–262. PMLR, 2017.

[5] Jean-Pierre Boeuf and Andrei Smolyakov. Physics and instabilities of low-temperature $E \times B$ plasmas for spacecraft propulsion and other applications. *Physics of Plasmas*, 30(5):050901, 2023.

[6] Peter G. Bolhuis, David Chandler, Christoph Dellago, and Phillip L. Geissler. Transition path sampling: Throwing ropes over rough mountain passes, in the dark. *Annual Review of Physical Chemistry*, 53(Volume 53, 2002):291–318, 2002.

[7] Nicolas Boumal. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023.

[8] Romain Brault, Markus Heinonen, and Florence Buc. Random fourier features for operator-valued kernels. In Robert J. Durrant and Kee-Eung Kim, editors, *Proceedings of The 8th Asian Conference on Machine Learning*, volume 63 of *Proceedings of Machine Learning Research*, pages 110–125, The University of Waikato, Hamilton, New Zealand, 16–18 Nov 2016. PMLR.

[9] Steven L. Brunton, Marko Budišić, Eurika Kaiser, and J. Nathan Kutz. Modern Koopman theory for dynamical systems. *SIAM Review*, 64(2):229–340, 2022.

[10] Rizwan Chaudhry and René Vidal. Initial-state invariant binet-cauchy kernels for the comparison of linear dynamical systems. In *52nd IEEE Conference on Decision and Control*, pages 5377–5384. IEEE, 2013.

[11] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 794–803. PMLR, 2018.

[12] Anoop Cherian and Suvrit Sra. Riemannian dictionary learning and sparse coding for positive definite matrices. *IEEE transactions on neural networks and learning systems*, 28(12):2859–2871, 2016.

[13] Reinart Coosemans, Wouter Dekeyser, and Martine Baelmans. Bayesian analysis of turbulent transport coefficients in 2d interchange dominated ExB turbulence involving flow shear. *Journal of Physics: Conference Series*, 1785:012001, 2021.

[14] Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In *International conference on machine learning*, pages 685–693. PMLR, 2014.

[15] Farbod Faraji, Maryam Reza, Aaron Knoll, and J. Nathan Kutz. Dynamic mode decomposition for data-driven analysis and reduced-order modelling of $E \times B$ plasmas: I. extraction of spatiotemporally coherent patterns. *Journal of Physics D: Applied Physics*, 57:065201, 2024.

[16] Tryphon T. Georgiou. Distances and riemannian metrics for spectral density functions. *IEEE Transactions on Signal Processing*, 55(8):3995–4003, 2007.

[17] Thibaut Germain, Rémi Flamary, Vladimir R. Kostic, and Karim Lounici. A spectral-grassmann wasserstein metric for operator representations of dynamical systems. 2026.

[18] P. Ghendrih, Y. Asahi, E. Caschera, G. Dif-Pradalier, P. Donnel, X. Garbet, C. Gillot, V. Grandgirard, G. Latu, Y. Sarazin, et al. Generation and dynamics of SOL corrugated profiles. *Journal of Physics: Conference Series*, 1125:012011, 2018.

[19] Philippe Ghendrih, Guilhem Dif-Pradalier, Olivier Panico, Yanick Sarazin, Hugo Bufferand, Guido Ciraolo, Peter Donnel, Nicolas Fedorczak, Xavier Garbet, Virginie Grandgirard, et al. Role of avalanche transport in competing drift wave and interchange turbulence. *Journal of Physics: Conference Series*, 2397:012018, 2022.

[20] Klaus Glashoff and Michael M Bronstein. Optimization on the biorthogonal manifold. *arXiv preprint arXiv:1609.04161*, 2016.

[21] GYSELAX Team. Tokam2D: A 2d spectral solver for turbulence schemes. `https://github.com/gyselax/tokam2d`, 2026. Accessed: 2026-05-01.

[22] Mehrtash Harandi, Conrad Sanderson, Chunhua Shen, and Brian C Lovell. Dictionary learning and sparse coding on grassmann manifolds: An extrinsic solution. In *Proceedings of the IEEE international conference on computer vision*, pages 3120–3127, 2013.

[23] Mehrtash T Harandi, Conrad Sanderson, Richard Hartley, and Brian C Lovell. Sparse coding and dictionary learning for symmetric positive definite matrices: A kernel approach. In *European conference on computer vision*, pages 216–229. Springer, 2012.

[24] Maximilian Herde, Bogdan Raonić, Tobias Rohner, Roger Käppeli, Roberto Molinaro, Emmanuel de Bézenac, and Siddhartha Mishra. Poseidon: Efficient foundation models for pdes, 2024.

[25] Jeffrey Ho, Yuchen Xie, and Baba Vemuri. On a nonlinear generalization of sparse coding and dictionary learning. In *International conference on machine learning*, pages 1480–1488. PMLR, 2013.

[26] Zhiwu Huang, Luc Van Gool, and Johan A. K. Suykens. Sparse coding and dictionary learning for linear dynamical systems. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 543–550, 2016.

[27] Issei Ishikawa, Keisuke Fujii, Masahiro Ikeda, Yuka Hashimoto, and Yoshinobu Kawahara. A metric on nonlinear dynamical systems with perron–frobenius operators. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[28] Hachem Kadri, Emmanuel Duflos, Philippe Preux, Stéphane Canu, Alain Rakotomamonjy, and Julien Audiffren. Operator-valued kernels for learning from functional response data. *Journal of Machine Learning Research*, 17(20):1–54, 2016.

[29] Yoshinobu Kawahara. Dynamic mode decomposition with reproducing kernels for koopman spectral analysis. *NeurIPS*, 2016.

[30] Stefan Klus, Feliks Nüske, Peter Koltai, Hao Wu, Ioannis G Kevrekidis, Christof Schütte, and Frank Noé. On the numerical approximation of the perron-frobenius and koopman operator. *Journal of Computational Dynamics*, 2018.

[31] Bernard O. Koopman. Hamiltonian systems and transformation in hilbert space. *Proceedings of the National Academy of Sciences*, 17(5):315–318, 1931.

[32] V. R. Kostic, P. Novelli, R. Grazzi, K. Lounici, and M. Pontil. Learning invariant representations of time-homogeneous stochastic dynamical systems. In *International Conference on Learning Representations*, 2024.

[33] Vladimir Kostic, Karim Lounici, Pietro Novelli, and Massimiliano Pontil. Sharp spectral rates for koopman operator learning. *Advances in Neural Information Processing Systems*, 36:32328–32339, 2023.

[34] Vladimir Kostic, Pietro Novelli, Andreas Maurer, Carlo Ciliberto, Lorenzo Rosasco, and Massimiliano Pontil. Learning dynamical systems via koopman operator regression in reproducing kernel hilbert spaces. *Advances in Neural Information Processing Systems*, 35:4017–4031, 2022.

[35] Vladimir R. Kostic, Karim Lounici, Hélène Halconruy, Timothée Devergne, Pietro Novelli, and Massimiliano Pontil. Laplace transform based low-complexity learning of continuous markov semigroups. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*. PMLR, 2025.

[36] Vladimir R. Kostic, Karim Lounici, Grégoire Pacreau, Giacomo Turri, Pietro Novelli, and Massimiliano Pontil. Neural conditional probability for uncertainty quantification. In *Advances in Neural Information Processing Systems*, 2024.

[37] Vladimir R. Kostic, Karim Lounici, and Massimiliano Pontil. Toeplitz based spectral methods for data-driven dynamical systems, 2026.

[38] Vladimir Kostić, Jean-Baptiste Fermanian, et al. Kernel methods for koopman operator learning. In *NeurIPS*, 2022.

[39] Samuel Lanthaler and Nicholas H. Nelsen. Error bounds for learning with vector-valued random features. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[40] Andrzej Lasota and Michael C. Mackey. *Chaos, Fractals, and Noise*, volume 97 of *Applied Mathematical Sciences*. Springer New York, 1994.

[41] Peihua Li, Qilong Wang, Wangmeng Zuo, and Lei Zhang. Log-euclidean kernels for sparse representation and dictionary learning. In *Proceedings of the IEEE international conference on computer vision*, pages 1601–1608, 2013.

[42] Qianxiao Li, Felix Dietrich, Erik M. Bollt, and Ioannis G. Kevrekidis. Extended dynamic mode decomposition with dictionary learning: a data-driven adaptive spectral decomposition of the koopman operator. *Chaos*, 27(10):103111, 2017.

[43] 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.

[44] Julien Mairal, Francis Bach, and Jean Ponce. Task-driven dictionary learning. *IEEE transactions on pattern analysis and machine intelligence*, 34(4):791–804, 2011.

[45] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th annual international conference on machine learning*, pages 689–696, 2009.

[46] Richard J Martin. A metric for arma processes. *IEEE transactions on Signal Processing*, 48(4):1164–1170, 2002.

[47] Igor Mezić. On comparison of dynamics of dissipative and finite-time systems using koopman operator methods. *IFAC-PapersOnLine*, 49(18):454–461, 2016.

[48] Igor Mezić and Andrzej Banaszuk. Comparison of systems with complex behavior. *Physica D: Nonlinear Phenomena*, 197(1-2):101–133, 2004.

[49] Hà Quang Minh. Operator-valued bochner theorem, Fourier feature maps for operator-valued kernels, and vector-valued learning. *arXiv preprint arXiv:1608.05639*, 2016.

[50] Bruno A Olshausen and David J Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

[51] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

[52] Emmanuel Rio. Covariance inequalities for strongly mixing processes. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 29(4):587–597, 1993.

[53] Sheldon M Ross. *Stochastic Processes*. John Wiley & Sons, 1995.

[54] Morgan A Schmitz, Matthieu Heitz, Nicolas Bonneel, Fred Ngole, David Coeurjolly, Marco Cuturi, Gabriel Peyré, and Jean-Luc Starck. Wasserstein dictionary learning: Optimal transport-based unsupervised nonlinear dictionary learning. *SIAM Journal on Imaging Sciences*, 11(1):643–678, 2018.

[55] Ch. Schütte, W. Huisinga, and P. Deuflhard. Transfer operator approach to conformational dynamics in biomolecular systems. In Bernold Fiedler, editor, *Ergodic Theory, Analysis, and Efficient Simulation of Dynamical Systems*, pages 191–223, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg.

[56] Christof Schütte, Stefan Klus, and Carsten Hartmann. Overcoming the timescale barrier in molecular dynamics: Transfer operators, variational principles and machine learning. *Acta Numerica*, 32:517–673, 2023.

[57] Roy Taylor, J. Nathan Kutz, Kyle D. Morgan, and Brian A. Nelson. Dynamic mode decomposition for plasma diagnostics and validation. *Review of Scientific Instruments*, 89(5):053501, 2018.

[58] Ivana Tošić and Pascal Frossard. Dictionary learning. *IEEE Signal Processing Magazine*, 28(2):27–38, 2011.

[59] Nilesh Tripuraneni, Chi Jin, and Michael I. Jordan. Provable meta-learning of linear representations. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[60] Hien Van Nguyen, Vishal M Patel, Nasser M Nasrabadi, and Rama Chellappa. Kernel dictionary learning. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2021–2024. IEEE, 2012.

[61] Hien Van Nguyen, Vishal M Patel, Nasser M Nasrabadi, and Rama Chellappa. Design of non-linear kernel dictionaries for object recognition. *IEEE Transactions on Image Processing*, 22(12):5123–5135, 2013.

[62] S. V. N. Vishwanathan, Alexander J. Smola, and René Vidal. Binet-cauchy kernels on dynamical systems and its application to the analysis of dynamic scenes. *International Journal of Computer Vision*, 73(1):95–119, 2007.

[63] Matthew O Williams, Clarence W Rowley, and Ioannis G Kevrekidis. A kernel-based method for data-driven koopman spectral analysis. *Journal of Computational Dynamics*, 2015.

[64] Hao Wu, Feliks Nüske, Fabian Paul, Stefan Klus, Péter Koltai, and Frank Noé. Variational Koopman models: Slow collective variables and molecular kinetics from short off-equilibrium simulations. *The Journal of Chemical Physics*, 146(15):154104, 2017.

A Operator-Theoretic Foundations

A.1 Operator representations and estimation

Markov semigroup and transfer operator. Let $(X_t)_{t\geq 0}$ be a time-homogeneous Markov process on a measurable state space $\mathcal{X}$, and let $\mathcal{F} \subset \mathcal{L}^2_\pi(\mathcal{X})$ be a Hilbert space of observables. The evolution of the system is described by the Markov semigroup $(A_t)_{t\geq 0}$ acting on $\mathcal{F}$, defined by

$$[A_t f](x) = \mathbb{E}[f(X_t) \mid X_0 = x], \quad f \in \mathcal{F}. \tag{9}$$

The family $(A_t)_{t\geq 0}$ is a strongly continuous semigroup satisfying

$$A_0 = \mathrm{Id}, \qquad A_{t+s} = A_t A_s, \quad \forall t, s \geq 0.$$

Infinitesimal generator. The infinitesimal generator G associated with (A_t) is defined as

$$Gf = \lim_{t\to 0^+} \frac{A_t f - f}{t}, \tag{10}$$

with domain $\mathcal{D}(G) \subset \mathcal{F}$. The generator is, in general, an unbounded linear operator on $\mathcal{F}$ and characterizes the semigroup through $A_t = \exp(tG)$.

Spectral decomposition via compact resolvent. The spectral analysis of G requires additional structure due to its infinite-dimensional and unbounded nature. Let π be an invariant measure of the process and consider the Hilbert space $\mathcal{L}^2_\pi(\mathcal{X})$.

We assume that G has compact resolvent, i.e., there exists $\mu_0 \in \rho(G)$ such that $(\mu_0 \,\mathrm{Id} - G)^{-1}$ is compact on $\mathcal{L}^2_\pi(\mathcal{X})$. Under this assumption, the spectrum of G is discrete and consists of isolated eigenvalues $(\lambda_i)_{i\in\mathbb{N}}$ with finite multiplicity and no accumulation point except possibly at infinity.

If, in addition, G is self-adjoint, there exists an orthonormal basis $(f_i)_{i\in\mathbb{N}}$ of $\mathcal{L}^2_\pi(\mathcal{X})$ such that

$$Gf_i = \lambda_i f_i. \tag{11}$$

For each isolated eigenvalue λ_i, the associated spectral projector P_i is defined as the orthogonal projector onto the corresponding eigenspace in $\mathcal{L}^2_\pi(\mathcal{X})$.

In the simple eigenvalue case, we have

$$P_i f = \langle f, f_i \rangle_{\mathcal{L}^2_\pi(\mathcal{X})} f_i. \tag{12}$$

More generally, if λ_i has multiplicity m_i and $(f_{i,k})_{k=1}^{m_i}$ is an orthonormal basis of the eigenspace, then

$$P_i f = \sum_{k=1}^{m_i} \langle f, f_{i,k} \rangle_{\mathcal{L}^2_\pi(\mathcal{X})} f_{i,k}. \tag{13}$$

The projectors satisfy

$$P_i P_j = \delta_{ij} P_i, \qquad \sum_i P_i = \mathrm{Id} \tag{14}$$

on the spectral subspace.

With this notation, the generator admits the spectral representation

$$Gf = \sum_i \lambda_i P_i f, \tag{15}$$

for all $f \in \mathcal{D}(G)$ for which the expansion is well-defined, and the semigroup diagonalizes as

$$A_t f = \sum_i e^{\lambda_i t} P_i f. \tag{16}$$

Spectral perturbation under RKHS estimation. The true spectral projectors are defined in the ambient space $\mathcal{L}^2_\pi(\mathcal{X})$, whereas data-driven estimators are learned on an RKHS $\mathcal{H}$ whose geometry generally differs from that of $\mathcal{L}^2_\pi(\mathcal{X})$. We therefore do not compare the generator G directly with its estimator in the $\mathcal{L}^2_\pi(\mathcal{X}) \to \mathcal{L}^2_\pi(\mathcal{X})$ operator norm. Instead, following the spectral perturbation framework of [33], we compare the resolvent of the generator after embedding the RKHS into the appropriate ambient space.

Let $\mu > 0$ and write

$$R_\mu := (\mu \,\mathrm{Id} - G)^{-1}. \tag{17}$$

Let $Z_\pi : \mathcal{H} \to \mathcal{L}^2_\pi(\mathcal{X})$ denote the canonical injection into $\mathcal{L}^2_\pi(\mathcal{X})$, and let

$$\widehat{R}_\mu : \mathcal{H} \to \mathcal{H} \tag{18}$$

be an estimator of the resolvent constructed from data. We measure the operator error by

$$\mathcal{E}(\widehat{R}_\mu) := \|R_\mu Z_\pi - Z_\pi \widehat{R}_\mu\|_{\mathcal{H} \to \mathcal{L}^2_\pi(\mathcal{X})}. \tag{19}$$

Since spectral perturbation is evaluated in the ambient geometry, one must also control the distortion between the RKHS norm and the ambient norm. For $h \in \mathcal{H}$, define

$$\mathfrak{d}(h) := \frac{\|h\|_{\mathcal{H}}}{\|Z_\pi h\|_{\mathcal{L}^2_\pi(\mathcal{X})}}. \tag{20}$$

Let $\widehat{R}_\mu$ admit spectral pairs

$$\widehat{R}_\mu \, \widehat{h}_i = \widehat{\rho}_i \widehat{h}_i, \tag{21}$$

where $\widehat{\rho}_i$ approximates $(\mu - \lambda_i)^{-1}$. After embedding and normalization, define

$$\widehat{f}_i := \frac{Z_\pi \widehat{h}_i}{\|Z_\pi \widehat{h}_i\|_{\mathcal{L}^2_\pi(\mathcal{X})}}. \tag{22}$$

If the corresponding eigenvalue $\rho_i = (\mu - \lambda_i)^{-1}$ of R_μ is isolated with spectral gap gap_i, then Davis–Kahan-type perturbation arguments yield

$$|\rho_i - \widehat{\rho}_i| \;\le\; \mathcal{E}(\widehat{R}_\mu)\, \mathfrak{d}(\widehat{h}_i), \tag{23}$$

and

$$\|\widehat{f}_i - f_i\|^2_{\mathcal{L}^2_\pi(\mathcal{X})} \;\le\; \frac{2\,\mathcal{E}(\widehat{R}_\mu)\, \mathfrak{d}(\widehat{h}_i)}{[\mathrm{gap}_i - \mathcal{E}(\widehat{R}_\mu)\, \mathfrak{d}(\widehat{h}_i)]_+}. \tag{24}$$

Finally, generator eigenvalues are recovered via

$$\widehat{\lambda}_i = \mu - \widehat{\rho}_i^{-1}. \tag{25}$$

Thus, spectral estimation is controlled by two factors: the embedded operator error $\mathcal{E}(\widehat{R}_\mu)$ and the metric distortion $\mathfrak{d}(\widehat{h}_i)$.

Learning initial generators via spectral filtering. The first step consists in learning a collection of initial generator atoms directly from trajectory data by estimating spectral components of the infinitesimal generator G. Rather than approximating G through finite differences of the form $(A_{\Delta t} - I)/\Delta t$, which is unstable at small time scales, we rely on spectral filtering based on Toeplitz representations of analytic functions of the generator.

Let $A_{\Delta t} = e^{\Delta t G}$ denote the transfer operator at time step Δt. We consider a Toeplitz symbol

$$T(z) = \sum_{j \in \mathbb{Z}} a_j z^j, \tag{26}$$

and define the associated filtered operator

$$F(G) = T(A_{\Delta t}). \tag{27}$$

Since $A_{j\Delta t} = A_{\Delta t}^j$, the operator $F(G)$ admits the representation

$$F(G) = \sum_{j\in\mathbb{Z}} a_j A_{\Delta t}^j, \tag{28}$$

which can be approximated from data using time-lagged observations. In practice, we use a truncated expansion

$$F_\ell(G) = \sum_{|j|\le\ell} a_j A_{\Delta t}^j, \tag{29}$$

whose empirical action is represented by a banded Toeplitz matrix acting on the time-ordered data sequence. Thus, analytic functional calculus of the generator reduces to structured linear algebra on trajectory data.

Resolvent and exponential filters. To probe the spectrum of the generator G, we use families of analytic filters.

The key object is the generator resolvent

$$R_\mu = (\mu I - G)^{-1} = \int_0^\infty e^{tG} e^{-\mu t}\, dt, \tag{30}$$

which we approximate by a Toeplitz-weighted sum of transfer operators,

$$R_{\mu,\ell} = \sum_{j=0}^{\ell} a_j A_{\Delta t}^j, \qquad a_j \approx \Delta t\, e^{-\mu j \Delta t}, \tag{31}$$

corresponding to a discretization of the Laplace transform.

Equivalently, using the transfer-operator resolvent, we have

$$(e^\mu I - A_{\Delta t})^{-1} = \sum_{j=0}^{\infty} e^{-(j+1)\mu} A_{\Delta t}^j, \tag{32}$$

which naturally yields Toeplitz coefficients. These filters concentrate spectral information near the shift parameter μ, allowing us to localize different regions of the spectrum.

In addition, exponential and trigonometric filters (e.g. $e^{\Delta t G}$, $\cosh(\Delta t G)$, $\sinh(\Delta t G)$) can be used to emphasize specific spectral structures depending on the nature of the dynamics.

Spectral representation learning. For each filter $F_q(G)$, indexed by a parameter q such as a resolvent shift μ, we do not first estimate $F_q(G)$ as an operator on the ambient space. Instead, we use the Toeplitz filter to learn a low-dimensional latent spectral representation.

Given the Toeplitz coefficients associated with F_q, we form the Toeplitz-weighted lagged data operator. In the finite-dimensional representation this gives a filtered matrix W_q, while in the sample representation it gives the corresponding Toeplitz-filtered Gram operator. The latent spectral space is then obtained from the leading solutions of the generalized eigenproblem

$$W_q W_q^\top v_{q,k} = \sigma_{q,k}^2 C_\gamma v_{q,k}, \qquad k = 1, \dots, r_q, \tag{33}$$

or equivalently from its dual version. We denote the resulting latent space by

$$\mathcal{U}_q = \operatorname{span}\{v_{q,1}, \dots, v_{q,r_q}\}. \tag{34}$$

Only after this latent space has been learned do we represent the filtered dynamics on it. The empirical compressed filter is

$$\widehat{F}_q^{(r)} = V_q^\top W_q V_q, \tag{35}$$

where $V_q = [v_{q,1}, \dots, v_{q,r_q}]$. Its spectral decomposition

$$\widehat{F}_q^{(r)} w_{q,k} = \widehat{\nu}_{q,k} w_{q,k} \tag{36}$$

yields the filtered spectral coordinates. The corresponding generator atom is obtained by inverting the scalar filter,

$$\widehat{\lambda}_{q,k} = F_q^{-1}(\widehat{\nu}_{q,k}). \tag{37}$$

For the resolvent filter $F_q(\lambda) = (\mu - \lambda)^{-1}$, this gives

$$\widehat{\lambda}_{q,k} = \mu - \widehat{\nu}_{q,k}^{-1}. \tag{38}$$

Thus, Step 1 learns generator atoms by first learning a latent spectral representation induced by the Toeplitz-filtered trajectory, and then diagonalizing the compressed filtered operator in that learned latent space.

Latent generator atoms. The outcome of this step is a collection of spectral atoms

$$\mathcal{A}_0 = \left\{ \left(\widehat{\lambda}_{q,k}, \widehat{\psi}_{q,k}, q\right) : q \in \mathcal{Q},\ k \le r_q \right\}, \tag{39}$$

which define a latent spectral representation of the generator. Each atom corresponds to a localized spectral component obtained through a specific filter q.

These atoms provide a structured, low-dimensional parametrization of the generator spectrum and its associated modes. They serve as the input for the second step, where dictionary learning is performed on the manifold generated by these initial spectral representations.

Related work on operator estimation. Estimating evolution operators from data is challenging, as dynamical systems are typically observed through discrete trajectories, and neither the generator G nor its domain is known. A common approach consists in learning the action of the time-δ_t operator $A = \exp(\delta_t G)$ on a predefined Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$, yielding an estimator of the projected operator $\tilde{A} = P_{\mathcal{H}} A_{|_{\mathcal{H}}}$. This is typically achieved via empirical risk minimization with low-rank constraints [29, 34], and is closely related to kernel-based variants of dynamic mode decomposition [9].

An alternative approach consists in learning finite-dimensional function spaces, for instance parameterized by neural networks, in which the operator is well approximated [43, 32]. These methods aim at minimizing projection errors of the form $\|P_{\mathcal{H}}^{\perp} A_{|_{\mathcal{H}}}\|$, thereby adapting the representation space to the underlying dynamics.

More recently, spectral approaches based on analytic functional calculus have been proposed, where one does not directly approximate the generator or the transfer operator, but rather bounded transforms such as resolvents or Laplace-type operators. In particular, Toeplitz-based spectral methods [37] leverage structured combinations of time-lagged observations to approximate analytic functions of the generator, enabling stable recovery of spectral quantities through finite-dimensional projections.

While such approaches provide consistent estimators under suitable assumptions, they typically operate in high-dimensional function spaces and exhibit nonparametric statistical rates, where n denotes the number of samples [33]. This can limit their reliability in low-data regimes. Moreover, the resulting spectral decompositions lie on non-linear manifolds, making subsequent comparison and learning tasks non-trivial.

Learning transfer operators in RKHS. Let $\mathcal{D} = (x_i, y_i)_{i=1}^n$ be sampled from a stationary Markov process, with $y_i \sim P(\cdot \mid x_i)$. Let $\phi : \mathcal{X} \to \mathcal{H}$ be the canonical feature map and define

$$\Phi_X := [\phi(x_1), \dots, \phi(x_n)], \qquad \Phi_Y := [\phi(y_1), \dots, \phi(y_n)].$$

The empirical covariance and cross-covariance operators are

$$\widehat{C}_x = \frac{1}{n} \Phi_X \Phi_X^*, \qquad \widehat{C}_{xy} = \frac{1}{n} \Phi_X \Phi_Y^*.$$

For $\lambda > 0$ and rank r, we estimate the transfer operator A by the regularized reduced-rank estimator

$$\widehat{A}_{r,\lambda} \in \operatorname*{arg\,min}_{\operatorname{rank}(A) \le r} \left\{ \frac{1}{n} \sum_{i=1}^{n} \|A^* \phi(x_i) - \phi(y_i)\|_{\mathcal{H}}^2 + \lambda \|A\|_{\mathrm{HS}}^2 \right\}. \tag{40}$$

Equivalently, we first form the ridge-stabilized regression operator

$$\widehat{A}_\lambda := (\widehat{C}_x + \lambda \,\mathrm{Id})^{-1} \widehat{C}_{xy}, \tag{41}$$

and then retain its leading r spectral components. The regularization parameter λ controls stability, while r controls the approximation rank.

Dual representation. In practice, the estimator is computed through Gram matrices

$$K_{XX} := \Phi_X^* \Phi_X, \qquad K_{YY} := \Phi_Y^* \Phi_Y, \qquad K_{XY} := \Phi_X^* \Phi_Y.$$

Define

$$M_\lambda := (K_{XX} + n\lambda I)^{-1/2} K_{XY} (K_{YY} + n\lambda I)^{-1/2}. \tag{42}$$

If

$$M_\lambda = \sum_{j \geq 1} \widehat{\sigma}_j \widehat{u}_j \widehat{v}_j^\top,$$

its rank-r truncation is

$$M_{\lambda,r} := \sum_{j=1}^{r} \widehat{\sigma}_j \widehat{u}_j \widehat{v}_j^\top.$$

The corresponding RKHS operator is

$$\widehat{A}_{r,\lambda} = \Phi_X (K_{XX} + n\lambda I)^{-1/2} M_{\lambda,r} (K_{YY} + n\lambda I)^{-1/2} \Phi_Y^*. \tag{43}$$

All computations are thus performed via kernel evaluations and an $n \times n$ SVD.

Error decomposition. Let $A^\star_{r,\lambda}$ denote the population counterpart. Then

$$\mathcal{E}(\widehat{A}_{r,\lambda}) := \|AZ_\pi - Z_\pi \widehat{A}_{r,\lambda}\|_{\mathcal{H} \to \mathcal{L}^2_\pi(\mathcal{X})} \tag{44}$$

admits the decomposition

$$\mathcal{E}(\widehat{A}_{r,\lambda}) \lesssim \underbrace{\|AZ_\pi - Z_\pi A^\star_{r,\lambda}\|_{\mathcal{H} \to \mathcal{L}^2_\pi(\mathcal{X})}}_{\text{approximation error}} + \underbrace{\|Z_\pi (A^\star_{r,\lambda} - \widehat{A}_{r,\lambda})\|_{\mathcal{H} \to \mathcal{L}^2_\pi(\mathcal{X})}}_{\text{statistical error}}. \tag{45}$$

Finite-rank approximation. Let

$$AZ_\pi = \sum_{j \geq 1} \sigma_j \, u_j \otimes v_j \tag{46}$$

be the singular value decomposition of the embedded transfer operator. Then

$$\inf_{\mathrm{rank}(B) \leq r} \|AZ_\pi - B\|_{\mathcal{H} \to \mathcal{L}^2_\pi(\mathcal{X})} = \sigma_{r+1}(AZ_\pi), \tag{47}$$

and the optimal approximation is obtained by truncation.

Under the regularity condition

$$C_{xy} C_{xy}^* \preceq a^2 C_x^{1+\alpha}, \qquad \alpha \in (0, 2], \tag{48}$$

and the spectral decay condition

$$\lambda_j(C_x) \leq b j^{-1/\beta}, \qquad \beta \in (0, 1], \tag{49}$$

the approximation error satisfies

$$\|AZ_\pi - Z_\pi A^\star_{r,\lambda}\|_{\mathcal{H} \to \mathcal{L}^2_\pi(\mathcal{X})} \lesssim \lambda^{\alpha/2} + \sigma_{r+1}(AZ_\pi). \tag{50}$$

Statistical error. With probability at least $1 - \delta$,

$$\|Z_\pi (A^\star_{r,\lambda} - \widehat{A}_{r,\lambda})\|_{\mathcal{H} \to \mathcal{L}^2_\pi(\mathcal{X})} \lesssim \lambda^{-\beta/2} n^{-1/2} \log(\delta^{-1}). \tag{51}$$

Balancing bias and variance yields

$$\lambda \asymp n^{-1/(\alpha+\beta)}, \tag{52}$$

and therefore

$$\mathcal{E}(\widehat{A}_{r,\lambda}) \lesssim \sigma_{r+1}(AZ_\pi) + n^{-\frac{\alpha}{2(\alpha+\beta)}} \log(\delta^{-1}). \tag{53}$$

Thus, the total error decomposes into a finite-rank approximation term controlled by the spectral decay of the embedded transfer operator, and a statistical estimation term governed by (α, β, n).

Random Fourier Features approximation. When the kernel k is shift-invariant, i.e. $k(x, y) = \kappa(x - y)$ for some positive definite function κ, Bochner's theorem guarantees the existence of a spectral measure Λ such that

$$k(x, y) = \mathbb{E}_{\omega\sim\Lambda}\left[e^{i\omega^\top(x-y)}\right] = \mathbb{E}_{\omega\sim\Lambda, b\sim\mathrm{Unif}([0,2\pi])}\left[z_\omega(x)z_\omega(y)\right], \tag{54}$$

where $z_\omega(x) := \sqrt{2}\cos(\omega^\top x + b)$ is a random feature map [51]. Given p i.i.d. draws $(\omega_k, b_k)_{k=1}^p$ from $\Lambda \times \mathrm{Unif}([0, 2\pi])$, the random Fourier feature (RFF) map is

$$\phi(x) := \frac{1}{\sqrt{p}}\left[z_{\omega_1}(x), \ldots, z_{\omega_p}(x)\right]^\top \in \mathbb{R}^p, \tag{55}$$

so that $\phi(x)^\top\phi(y) \approx k(x, y)$. The RFF approximation replaces the potentially infinite-dimensional RKHS $\mathcal{H}$ by the finite-dimensional feature space $\mathbb{R}^p$, with approximation error controlled by McDiarmid's inequality: with probability at least $1 - \delta$ over the draw of features,

$$\sup_{x,y\in\mathcal{X}} |k(x, y) - \phi(x)^\top\phi(y)| \;\le\; \sqrt{\frac{2\log(2/\delta)}{p}}. \tag{56}$$

RFF in the DOODL framework. In our setting, the RFF map provides a natural finite-dimensional feature space satisfying the boundedness assumption $\kappa := \sup_x \|\phi(x)\|_2 < \infty$. Indeed, since $|z_\omega(x)| \le \sqrt{2}$ for all x, ω, b, we have

$$\|\phi(x)\|_2 = \frac{1}{\sqrt{p}}\left(\sum_{k=1}^{p} z_{\omega_k}(x)^2\right)^{1/2} \le \sqrt{2}, \qquad \forall x \in \mathcal{X}, \tag{57}$$

so $\kappa = \sqrt{2}$ uniformly in p. This makes the RFF map particularly well-suited for our statistical guarantees: the feature bound κ is dimension-free and does not degrade as p grows.

The RFF estimator of the transfer operator is obtained by substituting ϕ for the kernel feature map in the RRR estimator (40), yielding a finite-dimensional linear regression problem of size $p \times p$. This construction is consistent with the broader framework of random features for operator-valued and vector-valued regression [28, 49, 8, 39], which establishes that RFF-based estimators recover the statistical rates of their kernel counterparts when the number of features p is chosen appropriately relative to the sample size. Concretely, with p growing polynomially in n, the additional approximation error introduced by replacing the kernel Gram matrices with their RFF counterparts does not affect the leading statistical rate of the estimator [4, 51].

Finally, in the DOODL framework, the same RFF map ϕ is shared across all d training systems and the test system, ensuring a common finite-dimensional observable space $\mathbb{R}^p$ as required by Assumption (A1).

Remark 1 (Bounded activations as an alternative)**.** *When ϕ is the last layer of a neural network with bounded activation (e.g.* $\arctan$*, sigmoid, or* $\tanh$*), the same bound $\kappa < \infty$ holds by construction. In this case all the statistical guarantees of Appendix F apply without modification, with κ determined by the activation range rather than the RFF construction.*

A.2 Geometry of operators via spectral optimal transport

We consider operators only through their finite-rank components. Let $\mathcal{H}$ be a separable Hilbert space and fix $r \in \mathbb{N}^*$. Denote by $\mathcal{S}_r(\mathcal{H})$ the set of non-defective operators of rank at most r. In the setting of this work, each operator is replaced by an element $G \in \mathcal{S}_r(\mathcal{H})$ capturing its dominant spectral modes. These operators are assumed to lie in a common r-dimensional subspace, either by construction or through the estimation procedure described above.

Spectral decomposition as measures over eigen-components. Let $G \in \mathcal{S}_r(\mathcal{H})$ admit the spectral decomposition

$$G = \sum_{i=1}^{\ell} \lambda_i P_i, \tag{58}$$

where $(\lambda_i)_{i\in[\ell]}$ are distinct eigenvalues, P_i are the associated spectral projectors, and $m_i := \operatorname{rank}(P_i)$ their multiplicities, with $m_{\mathrm{tot}} := \sum_{i=1}^{\ell} m_i \le r$.

We encode the spectral information of G as a discrete probability measure on $\mathbb{C} \times \mathcal{E}_r(\mathcal{H})$, where $\mathcal{E}_r(\mathcal{H}) := \{P \in \mathcal{S}_r(\mathcal{H}) : P^2 = P\}$ denotes the set of oblique projections, by

$$\mu(G) = \sum_{i=1}^{\ell} \frac{m_i}{m_{\mathrm{tot}}}\, \delta_{(\lambda_i, P_i)}. \tag{59}$$

Spectral Optimal Transport between Operators. Let $d_{\mathcal{E}}$ be a divergence on $\mathcal{E}_r(\mathcal{H})$. For $\eta \in (0,1)$ and $q \in \mathbb{N}^*$, define the ground cost

$$C_\eta^q\big((\lambda, P), (\lambda', P')\big) = \eta\, |\lambda - \lambda'|^q + (1-\eta)\, d_{\mathcal{E}}^q(P, P'). \tag{60}$$

Given $G, G' \in \mathcal{S}_r(\mathcal{H})$ with spectral measures

$$\mu(G) = \sum_{i=1}^{\ell} w_i\, \delta_{(\lambda_i, P_i)}, \qquad \mu(G') = \sum_{j=1}^{\ell'} w'_j\, \delta_{(\lambda'_j, P'_j)}, \tag{61}$$

their spectral optimal transport (SGOT) divergence is defined as

$$d_{\mathcal{S}}(G, G') = \min_{\pi \in \Pi(w, w')} \sum_{i=1}^{\ell} \sum_{j=1}^{\ell'} \pi_{ij}\, C_\eta^q\big((\lambda_i, P_i), (\lambda'_j, P'_j)\big), \tag{62}$$

where

$$\Pi(w, w') = \left\{ \pi \in \mathbb{R}_+^{\ell \times \ell'} : \sum_{j=1}^{\ell'} \pi_{ij} = w_i, \quad \sum_{i=1}^{\ell} \pi_{ij} = w'_j \right\}. \tag{63}$$

The divergence $d_{\mathcal{S}}$ compares operators through their spectral components by jointly aligning eigenvalues and spectral projectors. It is invariant under permutations of the spectral decomposition and depends only on the associated invariant subspaces. When $d_{\mathcal{E}}$ is a metric and $q = 1$, $d_{\mathcal{S}}$ coincides with a Wasserstein distance.

In the present work, G and G' are obtained as finite-rank estimators (or truncations) of underlying operators. Consequently, $\mu(G)$ and $\mu(G')$ are supported on finitely many atoms, and the computation of $d_{\mathcal{S}}$ reduces to a finite-dimensional optimal transport problem.

Note that $d_{\mathcal{S}}$ is in general a divergence, and reduces to a Wasserstein metric when $d_{\mathcal{E}}$ is a metric and $q = 1$. In this regime, $d_{\mathcal{S}}$ defines a well-posed objective for learning operators from data: given a collection of estimated operators, it can be used to fit a shared low-dimensional representation (or dictionary) by minimizing the discrepancy between their spectral measures, thereby aligning their dominant eigenvalues and invariant subspaces in a common latent space.

Metric between spectral projectors. Suppose two rank-one spectral projector in $\mathbb{C}^n$: $\mathbf{P} = \mathbf{u} \otimes \mathbf{v}$ and $\tilde{\mathbf{P}} = \tilde{\mathbf{u}} \otimes \tilde{\mathbf{v}}$. Let the spectral angle θ_s between $\mathbf{P}$ and $\tilde{\mathbf{P}}$ be defined as:

$$\cos(\theta_s) = \frac{|\langle \mathbf{P}, \tilde{\mathbf{P}} \rangle|}{\|\mathbf{P}\| \|\tilde{\mathbf{P}}\|} = \frac{|\langle \mathbf{u}, \tilde{\mathbf{u}} \rangle \langle \tilde{\mathbf{v}}, \mathbf{v} \rangle|}{\|\mathbf{u}\| \|\tilde{\mathbf{u}}\| \|\mathbf{v}\| \|\tilde{\mathbf{v}}\|}\,. \tag{64}$$

From spectral angle one can define metrics between rank-one operator detailed in Table 1.

Wasserstein barycenters of operators. Given a collection of finite-rank operators $(G^{(k)})_{k=1}^N \subset \mathcal{S}_r(\mathcal{H})$ with associated spectral measures $(\mu(G^{(k)}))_{k=1}^N$, a natural notion of average operator is provided by the Wasserstein barycenter of these measures under the SGOT geometry. For weights $(\alpha_k)_{k=1}^N$ with $\alpha_k \ge 0$ and $\sum_{k=1}^N \alpha_k = 1$, the barycenter is defined as

$$\bar{\mu} \in \operatorname*{arg\,min}_{\nu \in \mathcal{P}(\mathbb{C} \times \mathcal{E}_r(\mathcal{H}))} \sum_{k=1}^{N} \alpha_k\, W_{C_\eta^q}(\nu, \mu(G^{(k)})), \tag{65}$$

Table 1: Metrics between rank-one spectral projectors. $\cos(\theta_s) = \delta$.

Metric	Angle formulation	Matrix formulation
Geodesic (d_{geo})	θ_s^2	$\arccos^2(\delta)$
Chordal (d_{chord})	$\sin^2\theta_s$	$1-\delta^2$
Procrustes (d_{proc})	$2(1-\cos\theta_s)$	$2(1-\delta)$
Log-Martin (d_{Martin})	$-\log(\cos^2\theta_s)$	$-\log(\delta^2)$

where $W_{C_\eta^q}$ denotes the optimal transport cost induced by the ground metric C_η^q. When $q = 1$ and $d_{\mathcal{E}}$ is a metric, this corresponds to a Wasserstein barycenter in the space of spectral measures. In the finite-rank setting, $\bar{\mu}$ is supported on a finite number of atoms and can be written as

$$\bar{\mu} = \sum_{j=1}^{\bar{\ell}} \bar{w}_j \, \delta_{(\bar{\lambda}_j, \bar{P}_j)}, \tag{66}$$

which defines a barycentric operator

$$\bar{G} = \sum_{j=1}^{\bar{\ell}} \bar{\lambda}_j \bar{P}_j \in \mathcal{S}_r(\mathcal{H}). \tag{67}$$

This construction defines an operator-valued average that is consistent with the SGOT geometry: the barycenter $\bar{G}$ minimizes the total transport cost to the input operators, and therefore provides a representative element whose spectral measure is optimally aligned, in the sense of optimal transport, with those of the operators $(G^{(k)})_{k=1}^N$. In particular, the support of $\bar{\mu}$ captures spectral components that best match, under the joint eigenvalue–eigenspace cost, the dominant modes of the input operators. In our framework, such barycenters are used as representative elements (or dictionary atoms) for collections of estimated operators.

A.3 Dictionary learning on structured spaces

Dictionary learning as a reconstruction problem. Let $(x_i)_{i\in[N]}$ be data points in a space $\mathcal{X}$. Dictionary learning seeks a set of atoms $D = (D_\ell)_{\ell\in[K]} \subset \mathcal{X}$ and codes $(\alpha_i)_{i\in[N]}$, with $\alpha_i \in \Delta_K$, such that each data point is approximated through a decoding function B:

$$x_i \approx \mathrm{B}(\alpha_i; D). \tag{68}$$

The dictionary and codes are learned by minimizing a reconstruction objective

$$\min_{D,(\alpha_i)} \sum_{i=1}^{N} \mathcal{L}\big(x_i, \mathrm{B}(\alpha_i; D)\big) + \mathcal{R}(\alpha_i), \tag{69}$$

where $\mathcal{L}$ is a data fitting loss and $\mathcal{R}$ a regularization term.

A key component of the formulation in (69) is the decoding function B, which specifies how atoms are combined. In Euclidean settings, B is typically linear, $\mathrm{B}(\alpha; D) = \sum_\ell \alpha_\ell D_\ell$. More generally, when $\mathcal{X}$ is non-linear, B can be defined through barycentric combinations associated with a given geometry.

Dictionary learning was originally introduced in signal processing and computer vision with linear decoders and sparse regularization [50, 45], and has since been extended to a wide range of machine learning tasks [58]. Beyond Euclidean settings, several works have generalized this paradigm to structured spaces by adapting the decoding function B to the underlying geometry, including kernel-based methods [60, 61], Riemannian approaches on manifolds [25, 22], and formulations based on Wasserstein barycenters for probability measures [54]. These extensions highlight that dictionary learning fundamentally depends on the choice of geometry and associated notion of barycentric combination.

Limitations for operator-valued data. Several extensions of dictionary learning have been proposed for dynamical systems. In Koopman-based approaches, one may learn a dictionary of observables used to construct finite-dimensional approximations of a single operator, for instance within EDMD frameworks [42]. These methods learn the feature space on which an operator is represented, but do not learn a dictionary whose atoms are themselves dynamical operators, nor do they enable shared representations across systems.

Another line of work considers dictionary learning for linear dynamical systems (LDS), where both data points and atoms are parametric state-space models [26]. Such approaches rely on Euclidean operations on finite-dimensional parameterizations and are therefore tied to specific model classes.

These constructions do not extend to the setting considered in this work, where data points are dynamical systems represented by their associated operators. In this case, the relevant structure is spectral and operator-theoretic: eigenvalues and invariant subspaces encode the fundamental dynamical modes. Linear combinations in parameter space do not preserve this structure, and therefore do not yield meaningful interpolations between operators.

Operator dictionary learning via spectral optimal transport. In our setting, the data are finite-rank operators $(G^{(k)})_{k=1}^N \subset \mathcal{S}_r(\mathcal{H})$, each encoding the dominant spectral component of a dynamical system. We define dictionary learning directly in the SGOT geometry.

Let $\mu(G^{(k)})$ denote the spectral measure associated with $G^{(k)}$. Given $K \in \mathbb{N}^*$, we seek dictionary atoms $(D_\ell)_{\ell=1}^K \subset \mathcal{S}_r(\mathcal{H})$, with associated spectral measures $\mu(D_\ell)$. For a code $\alpha \in \Delta_K$, the reconstruction of an operator is defined through a barycentric combination in the SGOT geometry:

$$B_\alpha(D_1, \dots, D_K) \in \operatorname*{arg\,min}_{G \in \mathcal{S}_r(\mathcal{H})} \sum_{\ell=1}^K \alpha_\ell \, d_{\mathcal{S}}(G, D_\ell), \tag{70}$$

that is, B_α is a Wasserstein barycenter with respect to the SGOT divergence. Equivalently, at the level of spectral measures,

$$\mu(B_\alpha(D_1, \dots, D_K)) \in \operatorname*{arg\,min}_{\nu \in \mathcal{P}(\mathbb{C} \times \mathcal{E}_r(\mathcal{H}))} \sum_{\ell=1}^K \alpha_\ell \, d_{\mathcal{S}}\big(\nu, \mu(D_\ell)\big). \tag{71}$$

The operator dictionary learning problem is then given by

$$\min_{(D_\ell)_{\ell=1}^K \subset \mathcal{S}_r(\mathcal{H})} \min_{(\alpha^{(k)})_{k=1}^N \subset \Delta_K} \sum_{k=1}^N d_{\mathcal{S}} \left(G^{(k)}, B_{\alpha^{(k)}}(D_1, \dots, D_K) \right). \tag{72}$$

Thus, reconstruction is not performed by Euclidean superposition of operators, but by barycentric interpolation in the spectral optimal transport geometry. The learned atoms correspond to representative spectral structures, and the coefficients $\alpha^{(k)}$ provide low-dimensional coordinates describing each system relative to these prototypes.

B A Riemannian manifold of spectral decomposition

We provide a novel characterization of the manifold of spectral decompositions assuming low-rank settings and complex algebra. We also derive the tools required to perform Riemannian gradient-based optimization on this space.

B.1 Main results

Main results rely on preliminary results on manifolds of bi-orthogonal matrices which can be found in Section B.2.

The smooth manifold of spectral decompositions. Let $\mathcal{H}$ be a complex p-dimensional Hilbert space and $\mathcal{S}_r(\mathcal{H})$ being the set of non defective operators with rank at most r and simple spectrum. Any $G \in \mathcal{S}_r(\mathcal{H})$ admits a spectral decomposition with matrix representation:

$$\mathbf{G} = \mathbf{R} \operatorname{Diag}(\mathbf{\Lambda}) \mathbf{L}^* \quad \text{s.t.} \quad \mathbf{L}^* \mathbf{R} = \mathbf{I}_r \,, \tag{73}$$

where $\mathbf{\Lambda} \in \mathbb{C}^r$ are the eigenvalues, $(\mathbf{L}, \mathbf{R}) \in (\mathbb{C}^{p\times r})^2$ are the left/right eigenvectors satisfying the bi-orthogonal condition.

As depicted in Equation (77), the set of low-rank bi-orthogonal matrices $\mathcal{B} = \{(\mathbf{L}, \mathbf{R}) \in (\mathbb{C}^{p\times r})^2 \mid \mathbf{L}^*\mathbf{R} = \mathbf{I}_r\}$ forms a smooth manifold, hence by product of smooth manifold, the set of all possible spectral decomposition:

$$\mathcal{N} = \mathbb{C}^r \times \mathcal{B} = \{(\mathbf{\Lambda}, \mathbf{L}, \mathbf{R}) \mid \mathbf{L}^*\mathbf{R} = \mathbf{I}_r\}\,, \tag{74}$$

forms a smooth manifold.

However, operators' spectral decomposition are not unique: under the simple spectrum assumption, eigenvectors are defined up to independent rescaling. This induces a natural equivalence relation on spectral decompositions, corresponding to component-wise actions of $(\mathbb{C}^*)^r$ on eigenvectors, i.e. $(\mathbf{L}, \mathbf{R}) \sim (\mathbf{L}', \mathbf{R}')$ if and only if there exists $\mathbf{z} \in (\mathbb{C}^*)^r$ such that $(\mathbf{L}', \mathbf{R}') = (\mathbf{L}\,\mathrm{Diag}(\bar{\mathbf{z}}), \mathbf{R}\,\mathrm{Diag}(\mathbf{z}^{-1}))$. As depicted in lemma 2, this equivalence relationship defines smooth quotient manifold of bi-orthogonal phase/scale invariant matrices, denoted $\mathcal{B}/(\mathbb{C}^*)^r$. We therefore view spectral decompositions as elements of a quotient space

$$\mathcal{M} = \mathbb{C}^r \times \mathcal{B}/(\mathbb{C}^*)^r. \tag{75}$$

By product of quotient manifold, $\mathcal{M}$ is a smooth manifold, and $\mathcal{S}_r(\mathcal{H})$ can be identified with $\mathcal{M}$.

A Riemannian metric on the manifold of spectral decomposition. Since the quotient manifold $\mathcal{B}/(\mathbb{C}^*)^r$ can be endow with a Riemannian as described in Proposition 6 and Proposition 3, by product with an Euclidean space, the manifold of spectral decomposition $\mathcal{M}$, can be endowed with the Riemannian metric:

$$g_{(\mathbf{\Lambda},\mathbf{L},\mathbf{R})}((\boldsymbol{\lambda}, \boldsymbol{\xi}, \boldsymbol{\zeta}), (\boldsymbol{\lambda}', \boldsymbol{\xi}', \boldsymbol{\zeta}')) \triangleq \boldsymbol{\lambda}^*\boldsymbol{\lambda}' + \mathrm{Tr}(\boldsymbol{\xi}^*\boldsymbol{\xi}'(\mathbf{L}^*\mathbf{L})^{-1}) + \mathrm{Tr}(\boldsymbol{\zeta}^*\boldsymbol{\zeta}'(\mathbf{R}^*\mathbf{R})^{-1})\,, \tag{76}$$

where $(\mathbf{\Lambda}, \mathbf{L}, \mathbf{R}) \in \mathcal{M}$ and $(\boldsymbol{\lambda}, \boldsymbol{\xi}, \boldsymbol{\zeta}), (\boldsymbol{\lambda}', \boldsymbol{\xi}', \boldsymbol{\zeta}') \in T_{(\mathbf{\Lambda},\mathbf{L},\mathbf{R})}\mathcal{M}$ are tangent vectors. Since the part of the metric related to the eigenvalues is the natural inner product, the necessary tools to perform Riemannian gradient descent (retraction and Riemannian gradient and transport) are just composition of the indentity on the eigenvalue space with tools on $\mathcal{B}/(\mathbb{C}^*)^r$ defined in Section B.2.

B.2 Manifolds of bi-orthogonal low-rank matrices and subspaces

B.2.1 Smooth manifold structure of bi-orthogonal low-rank matrices and subspaces

Lemma 1 (Smooth manifold of bi-orthogonal low rank matrices.)**.** *Considering real differentiability, the manifold of low-rank bi-orthogonal complex matrices*

$$\mathcal{B} = \left\{(\mathbf{L}, \mathbf{R}) \in (\mathbb{C}^{p\times r})^2 \mid \mathbf{L}^*\mathbf{R} = \mathbf{I}_r\right\}\,, \tag{77}$$

with $r < p$, the manifold is a smooth manifold of dimension $2rp - r^2$ with tangent spaces:

$$\mathcal{T}_{(\mathbf{L},\mathbf{R})}\mathcal{B} = \left\{(\boldsymbol{\xi}, \boldsymbol{\zeta}) \in (\mathbb{C}^{p\times r})^2 \mid \boldsymbol{\xi}^*\mathbf{R} + \mathbf{L}^*\boldsymbol{\zeta} = \mathbf{0}_r\right\}, \quad \forall (\mathbf{L}, \mathbf{R}) \in \mathcal{B}\,. \tag{78}$$

Proof. Let consider the application:

$$\mathrm{F} : (\mathbf{L}, \mathbf{R}) \in (\mathbb{C}^{p\times r})^2 \mapsto \mathbf{L}^*\mathbf{R} - \mathbf{I}_r \in \mathbb{C}^{r\times r}\,. \tag{79}$$

As a quadratic function, it is a real-smooth map between two Hilbert spaces. In particular, $\mathcal{B} = \mathrm{F}^{-1}(\mathbf{0}_r)$ and since its differential (with Wirtinger notation):

$$\mathrm{DF}(\mathbf{L}, \mathbf{R})[\boldsymbol{\xi}, \boldsymbol{\zeta}] \triangleq \boldsymbol{\xi}^*\mathbf{R} + \mathbf{L}^*\boldsymbol{\zeta} \tag{80}$$

is surjective for any $(\mathbf{L}, \mathbf{R}) \in \mathrm{F}^{-1}(\mathbf{0}_r)$, by the preimage theorem, $\mathcal{B} = \mathrm{F}^{-1}(\mathbf{0}_r)$ is a smooth manifold with $\dim(\mathcal{B}) = 2nr - r^2$, and such that for any $(\mathbf{L}, \mathbf{R}) \in \mathrm{F}^{-1}(\mathbf{0}_r)$, its tangent space is defined as:

$$\mathcal{T}_{(\mathbf{L},\mathbf{R})}\mathcal{M} = \left\{(\boldsymbol{\xi}, \boldsymbol{\zeta}) \in (\mathbb{C}^{n\times r})^2 \mid \boldsymbol{\xi}^*\mathbf{R} + \mathbf{L}^*\boldsymbol{\zeta} = \mathbf{0}_r\right\}. \tag{81}$$

□

Retraction on $\mathcal{B}$. As $\mathcal{B}$ is submanifold of the Euclidean space $(\mathbb{C}^{p\times r})^2$, a retraction updates $(\mathbf{L},\mathbf{R})\in\mathcal{B}$ in the direction of a tangent vector $(\boldsymbol{\xi},\boldsymbol{\zeta})\in\mathcal{T}_{(\mathbf{L},\mathbf{R})}\mathcal{B}$ by projecting the vector $(\mathbf{L}+\boldsymbol{\xi},\mathbf{R}+\boldsymbol{\zeta})$ back on the manifold $\mathcal{B}$.

Proposition 1 (Retraction on $\mathcal{B}$)**.** *The application $R\triangleq R^L\circ R^R:\mathcal{TB}\mapsto\mathcal{B}$, such that for any $(\mathbf{L},\mathbf{R})\in\mathcal{B}$ and $(\boldsymbol{\xi},\boldsymbol{\zeta})\in\mathcal{T}_{(\mathbf{L},\mathbf{R})}\mathcal{B}$:*

$$R^R_{(\mathbf{L},\mathbf{R})}(\boldsymbol{\xi},\boldsymbol{\zeta})\triangleq(\boldsymbol{\xi},\mathbf{R}+\boldsymbol{\zeta})\ ,$$

$$R^L_{(\mathbf{L},\mathbf{R})}(\boldsymbol{\xi},\tilde{\mathbf{R}})\triangleq\left(\mathbf{L}+\boldsymbol{\xi}-\tilde{\mathbf{R}}(\tilde{\mathbf{R}}^*\tilde{\mathbf{R}})^{-1}(\tilde{\mathbf{R}}^*(\mathbf{L}+\boldsymbol{\xi})-\mathbf{I}_r),\tilde{\mathbf{R}}\right)\ ,$$

forms a retraction on $\mathcal{B}$.

Proof. Let $(\mathbf{L},\mathbf{R})\in\mathcal{B}$, since $\mathbf{L}^*\mathbf{R}=\mathbf{I}_r$, $\mathbf{L}$ and $\mathbf{R}$ are full-rank, hence by the Weyl inequality there exists a open neighborhood of $(\mathbf{0},\mathbf{0})$, denoted $\mathcal{U}$, in which $\mathbf{L}+\boldsymbol{\xi}$ and $\mathbf{R}+\boldsymbol{\zeta}$ remain full rank. Denoting $(\tilde{\mathbf{L}},\tilde{\mathbf{R}})=\mathrm{R}(\boldsymbol{\xi},\boldsymbol{\zeta})$, one can verify that $\tilde{\mathbf{L}}^*\tilde{\mathbf{R}}=\mathbf{I}_r$. Hence R is well defined on $\mathcal{U}$ and takes value in $\mathcal{B}$. In addition $\mathrm{R}(\mathbf{0},\mathbf{0})=(\mathbf{L},\mathbf{R})$. Since any $(\tilde{\mathbf{L}},\tilde{\mathbf{R}})\in\mathcal{U}$ are full rank, $\tilde{\mathbf{R}}^*\tilde{\mathbf{R}}$ is invertible, hence by addition, matrix product and inversion, the application R is smooth and verifies:

$$\mathrm{DR}_{(\mathbf{L},\mathbf{R})}(\mathbf{0},\mathbf{0})[\mathrm{d}\boldsymbol{\xi},\mathrm{d}\boldsymbol{\zeta}]\triangleq(\mathrm{d}\boldsymbol{\xi}-\mathbf{R}(\mathbf{R}^*\mathbf{R})^{-1}(\mathbf{R}^*\mathrm{d}\boldsymbol{\xi}+\mathrm{d}\boldsymbol{\zeta}^*\mathbf{L}),\mathrm{d}\boldsymbol{\zeta})\ . \tag{82}$$

However, since for any $(\mathrm{d}\boldsymbol{\xi},\mathrm{d}\boldsymbol{\zeta})\in\mathcal{T}_{(\mathbf{L},\mathbf{R})}\mathcal{M}$, $\mathbf{R}^*\mathrm{d}\boldsymbol{\xi}+\mathrm{d}\boldsymbol{\zeta}^*\mathbf{L}=\mathbf{0}$, it follows $\mathrm{DR}_{(\mathbf{L},\mathbf{R})}(\mathbf{0},\mathbf{0})=\mathrm{Id}$ on $\mathcal{U}$. Hence by definition 4.1.1 in [1], R is a retraction on $\mathcal{B}$. □

Remark 2. *Let $(\mathbf{L},\mathbf{R})\in\mathcal{B}$ and $(\boldsymbol{\xi},\boldsymbol{\zeta})\in\mathcal{T}_{(\mathbf{L},\mathbf{R})}\mathcal{B}$. To remain in the open subset where the retraction is well-defined, one can estimate $\lambda\in\mathbb{R}_+^*$ such that $\mathbf{R}+\lambda\boldsymbol{\zeta}$ is full rank, i.e. $\lambda<\sigma_{\min}(\mathbf{R})/\|\boldsymbol{\zeta}\|$, the norm being the operator norm or the HS norm.*

Since bi-orthogonal low-rank matrices from a smooth manifold, we can investigates properties of the manifold of phase/scale invariant bi-orthogonal matrices:

Lemma 2 (Smooth manifold of phase/scale invariant bi-orthogonal matrices.)**.** *The manifold $\mathcal{B}/(\mathbb{C}^*)^r$ with $(\mathbf{L},\mathbf{R})\sim(\mathbf{L}',\mathbf{R}')$ if and only if there exists $\mathbf{z}\in(\mathbb{C}^*)^r$ such that $(\mathbf{L}',\mathbf{R}')=(\mathbf{L}\,\mathrm{Diag}(\bar{\mathbf{z}}),\mathbf{R}\,\mathrm{Diag}(\mathbf{z}^{-1}))$, is a smooth quotient manifold and the quotient map $\pi:\mathcal{B}\mapsto\mathcal{B}/(\mathbb{C}^*)^r$ is a submersion.*

Proof. As $((\mathbb{C}^*)^r,\times)$ is an Abelian Lie group for the element-wise product, the action $(\mathbf{z},(\mathbf{L},\mathbf{R}))\triangleq(\mathbf{L}\,\mathrm{Diag}(\bar{\mathbf{z}}),\mathbf{R}\,\mathrm{Diag}(\mathbf{z}^{-1}))$, for any $(z,(\mathbf{L},\mathbf{R}))\in(\mathbb{C}^*)^r\times\mathcal{B}$ is well defined and smooth. For any $(\mathbf{L},\mathbf{R})\in\mathcal{B}$, the stabilizer set $G_{(\mathbf{L},\mathbf{R})}=\{z\in(\mathbb{C}^*)^r\mid(\mathbf{L},\mathbf{R})=\mathbf{z}\cdot(\mathbf{L},\mathbf{R})\}=\{\mathbf{I}_r\}$ is reduced to the neutral, hence the action is free. We prove the action's properness with the sequential characterization. Let $(\mathbf{L}_n,\mathbf{R}_n)\to(\mathbf{L},\mathbf{R})$ and $(\mathbf{L}_n\,\mathrm{Diag}(\bar{\mathbf{z}}_n),\mathbf{R}_n\,\mathrm{Diag}(\mathbf{z}_n^{-1}))\to(\mathbf{L}',\mathbf{R}')$, let's prove that $\mathbf{z}_n\to\mathbf{z}$. By convergence, since for $n\in\mathbb{N}$, $\mathbf{L}$ (resp. $\mathbf{R}$) is full rank, there exist constants $\mathrm{c}_{\mathbf{L}},\mathrm{C}_{\mathbf{L}}>0$ (resp. $\mathrm{c}_{\mathbf{R}},\mathrm{C}_{\mathbf{R}}$) such that for any $n\in\mathbb{N}$, $\mathrm{c}_{\mathbf{L}}\le\|\mathbf{L}_n\|\le\mathrm{C}_{\mathbf{L}}$ (resp. $\mathrm{c}_{\mathbf{R}}\le\|\mathbf{R}_n\|\le\mathrm{C}_{\mathbf{R}}$). Similarly, there exist $\mathrm{m}_{\mathbf{L}},\mathrm{M}_{\mathbf{L}}>0$ (resp. $\mathrm{m}_{\mathbf{R}},\mathrm{M}_{\mathbf{R}}$) such that for any $n\in\mathbb{N}$, $\mathrm{m}_{\mathbf{L}}\le\|\mathbf{L}_n\,\mathrm{Diag}(\bar{\mathbf{z}}_n)\|\le\mathrm{M}_{\mathbf{L}}$ (resp. $\mathrm{m}_{\mathbf{R}}\le\|\mathbf{R}_n\,\mathrm{Diag}(\mathbf{z}_n^{-1})\|\le\mathrm{M}_{\mathbf{R}}$). Hence, for any $n\in\mathbf{N}$:

$$\frac{\mathrm{m}_{\mathbf{L}}}{\mathrm{C}_{\mathbf{L}}}\le\frac{\|\mathbf{L}_n\,\mathrm{Diag}(\bar{\mathbf{z}}_n)\|}{\|\mathbf{L}_n\|}\le\|\mathbf{z}_n\|\quad\text{and}\quad\frac{\mathrm{m}_{\mathbf{R}}}{\mathrm{C}_{\mathbf{R}}}\le\frac{\|\mathbf{R}_n\,\mathrm{Diag}(\mathbf{z}_n^{-1})\|}{\|\mathbf{R}_n\|}\le\|\mathbf{z}_n^{-1}\|\ ,$$

leading to the bounds: $\mathrm{m}_{\mathbf{L}}/\mathrm{C}_{\mathbf{L}}\le\|\mathbf{z}_n\|\le r\mathrm{C}_{\mathbf{R}}/\mathrm{m}_{\mathbf{R}}$. As a sequence of in a compact set of $(\mathbb{C}^*)^r$, $(\mathbf{z}_n)_{n\in\mathbb{N}}$ admits a convergent subsequence. Hence, by the Bolzano-Weierstrass, the action is proper by compactness of the pre-image of a compact. Finally, as the action is smooth, free and proper, the quotient manifold is smooth and the quotient map $\pi:\mathcal{B}\mapsto\mathcal{B}/(\mathbb{C}^*)$ is a submersion. □

Remark 3. *The manifold $\mathcal{B}/(\mathbb{C}^*)^r$ also corresponds to the smooth manifold of set of r orthogonal one-dimensional subspace of $\mathbb{C}^p$.*

B.2.2 Riemannian metrics on the $\mathcal{B}$ and $\mathcal{B}/(\mathbb{C}^*)^r$

A natural Riemannian metric. Since $(\mathbb{C}^{p\times r})^2$ endowed with the inner product $\langle(\mathbf{A},\mathbf{U}),(\mathbf{B},\mathbf{V})\rangle \triangleq \mathrm{Tr}(\mathbf{A}^*\mathbf{B}) + \mathrm{Tr}(\mathbf{U}^*\mathbf{V})$ forms an Hilbert space, $\mathcal{B}$, as a submanifold, inherits a Riemannian metric g such that for any $(\boldsymbol{\xi},\boldsymbol{\zeta}),(\boldsymbol{\xi}',\boldsymbol{\zeta}') \in \mathcal{T}_{(\mathbf{L},\mathbf{R})}\mathcal{B}$,

$$g_{(\mathbf{L},\mathbf{R})}((\boldsymbol{\xi},\boldsymbol{\zeta}),(\boldsymbol{\xi}',\boldsymbol{\zeta}')) \triangleq \langle(\boldsymbol{\xi},\boldsymbol{\zeta}),(\boldsymbol{\xi}',\boldsymbol{\zeta}')\rangle \,. \tag{83}$$

In particular, for $\bar{f} : (\mathbb{C}^{p\times r})^2 \mapsto \mathbb{C}$, and f its restriction to $\mathcal{B}$ we have the relationship:

$$\mathrm{grad}_{(\mathbf{L},\mathbf{R})}\, f = \mathrm{P}_{(\mathbf{L},\mathbf{R})}(\mathrm{grad}_{(\mathbf{L},\mathbf{R})}\, \bar{f}) \,, \tag{84}$$

where $\mathrm{P}_{(\mathbf{L},\mathbf{R})} : (\mathbb{C}^{p\times r})^2 \mapsto \mathcal{T}_{(\mathbf{L},\mathbf{R})}\mathcal{B}$ is the orthogonal projector on the subspace $\mathcal{T}_{(\mathbf{L},\mathbf{R})}\mathcal{B}$ at $(\mathbf{L},\mathbf{R}) \in \mathcal{B}$.

Proposition 2 (Orthogonal projection on tangent spaces)**.** *For any $(\mathbf{L},\mathbf{R}) \in \mathcal{B}$ endowed with its natural Riemannian metric, the orthogonal projector on the tangent space $\mathcal{T}_{(\mathbf{L},\mathbf{R})}\mathcal{B}$ verifies:*

$$P_{(\mathbf{L},\mathbf{R})} : (\boldsymbol{\xi}',\boldsymbol{\zeta}') \in (\mathbb{C}^{p\times r})^2 \mapsto (\boldsymbol{\xi}' - \mathbf{R}\boldsymbol{\mu}^*, \boldsymbol{\zeta}' - \mathbf{L}\boldsymbol{\mu}) \in \mathcal{T}_{(\mathbf{L},\mathbf{R})}\mathcal{B} \,, \tag{85}$$

with:

$$\boldsymbol{\mu} = \mathbf{P}_{\mathbf{L}}\mathbf{Z}\mathbf{P}_{\mathbf{R}}^* \quad \textit{s.t.} \quad \mathbf{Z}_{ij} = \frac{[\mathbf{P}_{\mathbf{L}}^*(\boldsymbol{\xi}'^*\mathbf{R} + \mathbf{L}^*\boldsymbol{\zeta}')\mathbf{P}_{\mathbf{R}}]_{ij}}{\boldsymbol{\lambda}_{\mathbf{R},j} + \boldsymbol{\lambda}_{\mathbf{L},i}} \quad \textit{and} \quad \begin{cases} \mathbf{R}^*\mathbf{R} = \mathbf{P}_{\mathbf{R}}\,\mathrm{Diag}(\boldsymbol{\lambda}_{\mathbf{R}})\mathbf{P}_{\mathbf{R}}^* \\ \mathbf{L}^*\mathbf{L} = \mathbf{P}_{\mathbf{L}}\,\mathrm{Diag}(\boldsymbol{\lambda}_{\mathbf{L}})\mathbf{P}_{\mathbf{L}}^* \end{cases} . \tag{86}$$

Proof. Given $(\mathbf{L},\mathbf{R}) \in \mathcal{B}$, since $\mathcal{T}_{(\mathbf{L},\mathbf{R})}\mathcal{B}$ is a vector subspace of the Hilbert space $(\mathbb{C}^{p\times r})^2$, by the projection theorem on a closed convex set, the orthogonal projector exists, it is a linear application and it verifies:

$$\mathrm{P}_{(\mathbf{L},\mathbf{R})} : (\boldsymbol{\xi}',\boldsymbol{\zeta}') \in (\mathbb{C}^{p\times r})^2 \mapsto \underset{(\boldsymbol{\xi},\boldsymbol{\zeta})\in\mathcal{T}_{(\mathbf{L},\mathbf{R})}\mathcal{B}}{\arg\min} \left\|(\boldsymbol{\xi},\boldsymbol{\zeta}) - (\boldsymbol{\xi}',\boldsymbol{\zeta}')\right\|^2 . \tag{87}$$

Given $(\boldsymbol{\xi}',\boldsymbol{\zeta}') \in (\mathbb{C}^{p\times r})^2$ and considering the real differentiability case, the minimization problem is strictly convex and strictly feasible, hemce the KKT conditions are necessary and sufficient conditions to characterize the optimum. The Lagrandian can be expressed as:

$$\mathcal{L}(\boldsymbol{\xi},\boldsymbol{\zeta},\boldsymbol{\lambda},\tilde{\boldsymbol{\lambda}}) = \left\|(\boldsymbol{\xi},\boldsymbol{\zeta}) - (\boldsymbol{\xi}',\boldsymbol{\zeta}')\right\|^2 + \mathrm{Tr}(\boldsymbol{\lambda}^\intercal\,\mathrm{Re}(\boldsymbol{\xi}^*\mathbf{R} + \mathbf{L}^*\boldsymbol{\zeta})) + \mathrm{Tr}(\tilde{\boldsymbol{\lambda}}^\intercal\,\mathrm{Im}(\boldsymbol{\xi}^*\mathbf{R} + \mathbf{L}^*\boldsymbol{\zeta})) \tag{88}$$

As a quadratic function, the lagrangian is differentiable and, taking the Wirtinger notation, at the optimum it holds:

$$\begin{cases} \nabla_{\bar{\boldsymbol{\xi}}}\mathcal{L} = \boldsymbol{\xi} - \boldsymbol{\xi}' + \mathbf{R}(\boldsymbol{\lambda} - i\tilde{\boldsymbol{\lambda}})^\intercal = \mathbf{0} \\ \nabla_{\bar{\boldsymbol{\zeta}}}\mathcal{L} = \boldsymbol{\zeta} - \boldsymbol{\zeta}' + \mathbf{L}(\boldsymbol{\lambda} + i\tilde{\boldsymbol{\lambda}}) = \mathbf{0} \\ \nabla_{\boldsymbol{\lambda}}\mathcal{L} = \mathrm{Re}(\boldsymbol{\xi}^*\mathbf{R} + \mathbf{L}^*\boldsymbol{\zeta}) = \mathbf{0} \\ \nabla_{\tilde{\boldsymbol{\lambda}}}\mathcal{L} = \mathrm{Im}(\boldsymbol{\xi}^*\mathbf{R} + \mathbf{L}^*\boldsymbol{\zeta}) = \mathbf{0} \end{cases} \tag{89}$$

Denoting $\boldsymbol{\mu} = \boldsymbol{\lambda} + i\tilde{\boldsymbol{\lambda}}$, the optimal solution exists, is unique and verifies:

$$\begin{cases} \boldsymbol{\xi} = \boldsymbol{\xi}' - \mathbf{R}\boldsymbol{\mu}^* \\ \boldsymbol{\zeta} = \boldsymbol{\zeta}' - \mathbf{L}\boldsymbol{\mu} \end{cases} \quad \text{s.t.} \quad \boldsymbol{\mu}\mathbf{R}^*\mathbf{R} + \mathbf{L}^*\mathbf{L}\boldsymbol{\mu} = \boldsymbol{\xi}'^*\mathbf{R} + \mathbf{L}^*\boldsymbol{\zeta}' \tag{90}$$

Note that $\boldsymbol{\mu}$ is solution of a Sylvester system and since $\mathbf{L}$ and $\mathbf{R}$ are full rank, the matrices $\mathbf{R}^*\mathbf{R}$ and $\mathbf{L}^*\mathbf{L}$ are invertible and self adjoint, hence they admit a spectral decomposition with all eigenvalues real and strictly positive. Hence, $\boldsymbol{\mu}$ verifies:

$$\boldsymbol{\mu} = \mathbf{P}_{\mathbf{L}}\mathbf{Z}\mathbf{P}_{\mathbf{R}}^* \quad \text{s.t.} \quad \mathbf{Z}_{ij} = \frac{[\mathbf{P}_{\mathbf{L}}^*(\boldsymbol{\xi}'^*\mathbf{R} + \mathbf{L}^*\boldsymbol{\zeta}')\mathbf{P}_{\mathbf{R}}]_{ij}}{\boldsymbol{\lambda}_{\mathbf{R},j} + \boldsymbol{\lambda}_{\mathbf{L},i}} \quad \text{and} \quad \begin{cases} \mathbf{R}^*\mathbf{R} = \mathbf{P}_{\mathbf{R}}\,\mathrm{Diag}(\boldsymbol{\lambda}_{\mathbf{R}})\mathbf{P}_{\mathbf{R}}^* \\ \mathbf{L}^*\mathbf{L} = \mathbf{P}_{\mathbf{L}}\,\mathrm{Diag}(\boldsymbol{\lambda}_{\mathbf{L}})\mathbf{P}_{\mathbf{L}}^* \end{cases} . \tag{91}$$

□

A stable Riemannian metric. We propose a Riemannian metric that is numerically more stable by rescaling search directions and avoiding directions towards singularities. This metric has the benefit to also be well-defined on the quotient manifold $\mathcal{B}/(\mathbb{C}^*)^r$ as depicted below.

Proposition 3 (Stable Riemannian metric)**.** *For any $(\boldsymbol{\xi}, \boldsymbol{\zeta}), (\boldsymbol{\xi}', \boldsymbol{\zeta}') \in \mathcal{T}_{(\mathbf{L},\mathbf{R})}\mathcal{B}$, the application:*

$$g_{(\mathbf{L},\mathbf{R})}((\boldsymbol{\xi}, \boldsymbol{\zeta}), (\boldsymbol{\xi}', \boldsymbol{\zeta}')) \triangleq \mathrm{Tr}(\boldsymbol{\xi}^*\boldsymbol{\xi}'(\mathbf{L}^*\mathbf{L})^{-1}) + \mathrm{Tr}(\boldsymbol{\zeta}^*\boldsymbol{\zeta}'(\mathbf{R}^*\mathbf{R})^{-1}) \,, \tag{92}$$

defines a Riemannian metric on $\mathcal{B}$.

Proof. Let $(\mathbf{L}, \mathbf{R}) \in \mathcal{B}$, since $\mathbf{L}^*\mathbf{R} = \mathbf{I}_r$, $\mathbf{L}$ and $\mathbf{R}$ are full rank, leading to $\mathbf{L}^*\mathbf{L}$ and $\mathbf{R}^*\mathbf{R}$ being hermitian and positive definite. Thus, by sum of trace norms, $g_{(\mathbf{L},\mathbf{R})}$ is hermitian, positive definite (separated) and verifies the triangle inequality. Finally, for any smooth vector fields on $\mathcal{B}$, $\mathbf{X}, \mathbf{Y} \in \mathfrak{F}(\mathcal{B})$, the application:

$$(\mathbf{L}, \mathbf{R}) \in \mathcal{B} \mapsto g_{(\mathbf{L},\mathbf{R})}((\mathbf{X}^{\boldsymbol{\xi}}_{(\mathbf{L},\mathbf{R})}, \mathbf{X}^{\boldsymbol{\zeta}}_{(\mathbf{L},\mathbf{R})}), (\mathbf{Y}^{\boldsymbol{\xi}}_{(\mathbf{L},\mathbf{R})}, \mathbf{Y}^{\boldsymbol{\zeta}}_{(\mathbf{L},\mathbf{R})})) \,,$$

is smooth as trace of matrix products of smooth matrices. Hence g is a Riemannian metric. □

Proposition 4 (Orthogonal projection on tangent spaces)**.** *Let $(\mathbf{L}, \mathbf{R}) \in \mathcal{B}$ and consider the Hilbert space $((\mathbb{C}^{p\times r})^2, g_{(\mathbf{L},\mathbf{R})})$, the orthogonal projector on the tangent space $\mathcal{T}_{(\mathbf{L},\mathbf{R})}\mathcal{B}$ verifies:*

$$P_{(\mathbf{L},\mathbf{R})} : (\boldsymbol{\xi}', \boldsymbol{\zeta}') \in (\mathbb{C}^{p\times r})^2 \mapsto \begin{pmatrix} \boldsymbol{\xi}' - \mathbf{R}(\mathbf{R}^*\mathbf{R})^{-1}(\boldsymbol{\xi}'^*\mathbf{R} + \mathbf{L}^*\boldsymbol{\zeta}')^* \\ \boldsymbol{\zeta}' - \mathbf{L}(\mathbf{L}^*\mathbf{L})^{-1}(\boldsymbol{\xi}'^*\mathbf{R} + \mathbf{L}^*\boldsymbol{\zeta}') \end{pmatrix} \in \mathcal{T}_{(\mathbf{L},\mathbf{R})}\mathcal{B} \,, \tag{93}$$

Proof. Given $(\mathbf{L}, \mathbf{R}) \in \mathcal{B}$, since $(\mathcal{T}_{(\mathbf{L},\mathbf{R})}\mathcal{B}, g_{(\mathbf{L},\mathbf{R})})$ is a vector subspace of the Hilbert space $((\mathbb{C}^{p\times r})^2, g_{(\mathbf{L},\mathbf{R})})$, by the projection theorem on a closed convex set, the orthogonal projector exists, it is a linear application and it verifies:

$$\mathrm{P}_{(\mathbf{L},\mathbf{R})} : (\boldsymbol{\xi}', \boldsymbol{\zeta}') \in (\mathbb{C}^{p\times r})^2 \mapsto \underset{(\boldsymbol{\xi},\boldsymbol{\zeta})\in\mathcal{T}_{(\mathbf{L},\mathbf{R})}\mathcal{B}}{\arg\min} g^2_{(\mathbf{L},\mathbf{R})}((\boldsymbol{\xi} - \boldsymbol{\xi}', \boldsymbol{\zeta} - \boldsymbol{\zeta}'), (\boldsymbol{\xi} - \boldsymbol{\xi}', \boldsymbol{\zeta} - \boldsymbol{\zeta}')). \tag{94}$$

Given $(\boldsymbol{\xi}', \boldsymbol{\zeta}') \in (\mathbb{C}^{p\times r})^2$ and considering the real differentiability case, the minimization problem is strictly convex and strictly feasible, hence the KKT conditions are necessary and sufficient conditions to characterize the optimum. The Lagrangian can be expressed as:

$$\mathcal{L}(\boldsymbol{\xi}, \boldsymbol{\zeta}, \boldsymbol{\lambda}, \tilde{\boldsymbol{\lambda}}) = g^2_{(\mathbf{L},\mathbf{R})}((\boldsymbol{\xi}-\boldsymbol{\xi}', \boldsymbol{\zeta}-\boldsymbol{\zeta}'), (\boldsymbol{\xi}-\boldsymbol{\xi}', \boldsymbol{\zeta}-\boldsymbol{\zeta}')) + \mathrm{Tr}(\boldsymbol{\lambda}^\intercal \,\mathrm{Re}(\boldsymbol{\xi}^*\mathbf{R}+\mathbf{L}^*\boldsymbol{\zeta})) + \mathrm{Tr}(\tilde{\boldsymbol{\lambda}}^\intercal \,\mathrm{Im}(\boldsymbol{\xi}^*\mathbf{R}+\mathbf{L}^*\boldsymbol{\zeta})) \tag{95}$$

As a quadratic function, the Lagrangian is differentiable and, taking the Wirtinger notation, at the optimum it holds:

$$\begin{cases} \nabla_{\bar{\boldsymbol{\xi}}}\mathcal{L} = (\boldsymbol{\xi} - \boldsymbol{\xi}')(\mathbf{L}^*\mathbf{L})^{-1} + \mathbf{R}(\boldsymbol{\lambda} - i\tilde{\boldsymbol{\lambda}})^\intercal = \mathbf{0} \\ \nabla_{\bar{\boldsymbol{\zeta}}}\mathcal{L} = (\boldsymbol{\zeta} - \boldsymbol{\zeta}')(\mathbf{R}^*\mathbf{R})^{-1} + \mathbf{L}(\boldsymbol{\lambda} + i\tilde{\boldsymbol{\lambda}}) = \mathbf{0} \\ \nabla_{\boldsymbol{\lambda}}\mathcal{L} = \mathrm{Re}(\boldsymbol{\xi}^*\mathbf{R} + \mathbf{L}^*\boldsymbol{\zeta}) = \mathbf{0} \\ \nabla_{\tilde{\boldsymbol{\lambda}}}\mathcal{L} = \mathrm{Im}(\boldsymbol{\xi}^*\mathbf{R} + \mathbf{L}^*\boldsymbol{\zeta}) = \mathbf{0} \end{cases} \tag{96}$$

Denoting $\boldsymbol{\mu} = \boldsymbol{\lambda} + i\tilde{\boldsymbol{\lambda}}$, the optimal solution exists, is unique and verifies:

$$\begin{cases} \boldsymbol{\xi} = \boldsymbol{\xi}' - \mathbf{R}\boldsymbol{\mu}^*(\mathbf{L}^*\mathbf{L}) \\ \boldsymbol{\zeta} = \boldsymbol{\zeta}' - \mathbf{L}\boldsymbol{\mu}(\mathbf{R}^*\mathbf{R}) \end{cases} \quad \text{s.t.} \quad \boldsymbol{\xi}^*\mathbf{R} + \mathbf{L}^*\boldsymbol{\zeta} = 0 \tag{97}$$

□

Proposition 5 (Riemannian gradient)**.** *Consider $\bar{f} : (\mathbb{C}^{n\times r})^2 \mapsto \mathbb{R}$ a smooth function and let f denotes its restriction to $\mathcal{B}$, for any $(\mathbf{L}, \mathbf{R}) \in \mathcal{B}$:*

$$\mathrm{grad}\, f(\mathbf{L}, \mathbf{R}) \triangleq P_{(\mathbf{L},\mathbf{R})}(\nabla_{\mathbf{L}}\bar{f}(\mathbf{L}^*\mathbf{L}), \nabla_{\mathbf{R}}\bar{f}(\mathbf{R}^*\mathbf{R})) \,, \tag{98}$$

where $\nabla\bar{f} = (\nabla_{\mathbf{L}}\bar{f}, \nabla_{\mathbf{R}}\bar{f})$ is the gradient in the natural Euclidean space and $P_{(\mathbf{L},\mathbf{R})}$ the orthogonal projector on the tangent space at $(\mathbf{L}, \mathbf{R}) \in \mathcal{B}$.

Proof. Consider $(\mathbf{L}, \mathbf{R}) \in \mathcal{B}$ and let $\nabla\bar{f} = (\nabla_{\mathbf{L}}\bar{f}, \nabla_{\mathbf{R}}\bar{f})$ be the gradient of $\bar{f}$ in the natural Euclidean space $((\mathbb{C}^{n\times r})^2, \langle ., . \rangle)$. Then its gradients in $((\mathbb{C}^{n\times r})^2, g_{(\mathbf{L},\mathbf{R})})$, denoted as $\nabla\bar{f}_g$, verifies:

$$\langle \nabla\bar{f}, \boldsymbol{\xi} \rangle = g_{(\mathbf{L},\mathbf{R})}(\nabla\bar{f}_g, \boldsymbol{\xi}) \,, \quad \forall \boldsymbol{\xi} \in \mathcal{T}_{(\mathbf{L},\mathbf{R})}\mathcal{B} \,.$$

Hence, $(\nabla_{\mathbf{L}}\bar{f}_g, \nabla_{\mathbf{R}}\bar{f}_g) = (\nabla_{\mathbf{L}}\bar{f}(\mathbf{L}^*\mathbf{L}), \nabla_{\mathbf{R}}\bar{f}(\mathbf{R}^*\mathbf{R}))$, and, by projection on $\mathcal{T}_{(\mathbf{L},\mathbf{R})}\mathcal{B}$, the Riemannian gradient of f at $(\mathbf{L}, \mathbf{R})$, denoted $\mathrm{grad}\, f(\mathbf{L}, \mathbf{R})$, verifies eq. (98). □

With the sable metric in mind we can now define a Riemannian structure on the manifold of phase/scale invariant low-rank matrices.

Proposition 6. *The manifold $(\mathcal{B}/(\mathbb{C}^*)^r, g)$ is a Riemannian quotient manifold of $(\mathcal{B}, g)$ where g is the stable Riemannian metric.*

Proof. Let $(\mathbf{L}, \mathbf{R}) \in \mathcal{B}$, $(\boldsymbol{\xi}, \boldsymbol{\zeta})$ and $(\boldsymbol{\xi}', \boldsymbol{\zeta}') \in \mathcal{T}_{(\mathbf{L},\mathbf{R})}\mathcal{B}$, by the group action of $(\mathbb{C}^*)^r$, for any $\mathbf{z} \in (\mathbb{C}^*)^r$ it follows:

$$\begin{aligned} g_{\mathbf{z}\cdot(\mathbf{L},\mathbf{R})}(\mathrm{d}\mathbf{z}(\boldsymbol{\xi},\boldsymbol{\zeta}), \mathrm{d}\mathbf{z}(\boldsymbol{\xi}',\boldsymbol{\zeta}')) &= g_{(\bar{\mathbf{z}}\mathbf{L},\mathbf{z}^{-1}\mathbf{R})}((\bar{\mathbf{z}}\boldsymbol{\xi}, \mathbf{z}^{-1}\boldsymbol{\zeta}), (\bar{\mathbf{z}}\boldsymbol{\xi}', \mathbf{z}^{-1}\boldsymbol{\zeta}')) \\ &= \mathrm{Tr}(\mathrm{Diag}(|\mathbf{z}|^2)\boldsymbol{\xi}^*\boldsymbol{\xi}'(\mathbf{L}^*\mathbf{L})^{-1}\,\mathrm{Diag}(|\mathbf{z}|^{-2})) \\ &\quad + \mathrm{Tr}(\mathrm{Diag}(|\mathbf{z}|^{-2})\boldsymbol{\zeta}^*\boldsymbol{\zeta}'(\mathbf{R}^*\mathbf{R})^{-1}\,\mathrm{Diag}(|\mathbf{z}|^2)) \\ &= g_{(\mathbf{L},\mathbf{R})}((\boldsymbol{\xi},\boldsymbol{\zeta}),(\boldsymbol{\xi}',\boldsymbol{\zeta}'))\,, \end{aligned}$$

where $|\mathbf{z}|$ is the element with norm vector. Thus, g is invariant to the group action. Thus, $(\mathcal{B}/(\mathbb{C}^*)^r, g)$ is a Riemannian quotient manifold of $\mathcal{B}$. □

Remark 4. *For any smooth function $\bar{f} : \mathcal{B} \to \mathbb{R}$ also defining a function on the quotient manifold $f : \mathcal{B}/(\mathbb{C}^*)^r \to \mathbb{R}$, the unique gradient represent at $[x]$ in horizontal space at $\mathcal{T}_x\mathcal{B}$, denoted with abuse of notation* $\mathrm{grad}\, f([x])$*, verifies:* $\mathrm{grad}\, f([x]) = \mathrm{grad}\, \bar{f}(x)$.

C Dynamical Operator Dictionary Learning

C.1 Wasserstein reconstruction model

Since the data fitting loss function corresponds to a divergence, a natural choice of reconstruction map is the Fréchet mean barycenter with

$$\mathrm{B}^{\mathrm{OT}}(\boldsymbol{\alpha}; \overline{\boldsymbol{\mathcal{G}}}) \triangleq \underset{\boldsymbol{\mathcal{G}} \in \mathcal{N}}{\arg\min} \textstyle\sum_{j\in[d]} \alpha_j d_{\mathcal{S}}(\boldsymbol{\mathcal{G}}, \overline{\boldsymbol{\mathcal{G}}}_j)\,. \tag{99}$$

This reconstruction map is consistent with the geometry of spectral operators induced by the divergence $d_{\mathcal{S}}$ [17]. A similar strategy was proposed in [54] for performing non-linear DL is the space of probability distributions endowed with the Wasserstein metric. It corresponds to a free-support optimal transport barycenter [14], but classical fixed point iteration cannot be applied because of the Riemannian structure of the manifold. Instead, the barycenter can be computed via Riemannian gradient-based optimization as detailed in Appendix B. While this approach faithfully captures the underlying geometry, it incurs a significant computational cost due to the inner optimization problem, which can become prohibitive in practice.

C.2 Optimization scheme

We adopt a stochastic inexact block-coordinate strategy similar to the one proposed in [45] for euclidean spaces. Given a batch of size b, we perform two successive steps:

1. *Coefficient estimation:* fix the dictionary and optimize $\{\boldsymbol{\alpha}_i\}_{i\in[b]}$ via gradient descent (with softmax parametrization). Since the problem is not convex, we use an initialization of based on proximity to dictionary atoms, i.e. $\boldsymbol{\alpha}_i \propto \{-d_{\mathcal{S}}(\boldsymbol{\mathcal{G}}_i, \overline{\boldsymbol{\mathcal{G}}}_j)\}_{j\in[d]}$ to start the optimization in a relevant region of the parameter space. In practice we use the Adam optimizer with a learning rate of 0.5 and perform 100 iterations. Algorithm 1 describes the pseudo code.
2. *Dictionary update:* fix the coordinates $\{\boldsymbol{\alpha}_i\}_{i\in[b]}$ and update the dictionary with a Riemannian gradient step on $\mathcal{N}^d$ from the objective on the batch. We leverage the envelope theorem to ignore implicit gradients through $\{\boldsymbol{\alpha}_i\}_{i\in[b]}$. Default gradient steps has a learning rate of 1e-2.

Algorithm 2 describes the overall optimization procedure. Computation of the data fitting loss that is the SGOT divergence [17] requires the computation of optimal transport plans that we implemented in order to leverage the computational efficiency of GPUs.

Algorithm 1 Operator coefficient estimation

Require: $\boldsymbol{\mathcal{G}} \triangleq \{\boldsymbol{\mathcal{G}}_i\}_{i\in[N]} \in \mathcal{M}^N, \overline{\boldsymbol{\mathcal{G}}} \triangleq \{\overline{\boldsymbol{\mathcal{G}}}_i\}_{i\in[d]} \in \mathcal{M}^d$ ▷ *Operators / Dictionary*

$\boldsymbol{\alpha} \leftarrow \text{InitializationCoefficientLogit}(\boldsymbol{\mathcal{G}}, \overline{\boldsymbol{\mathcal{G}}})$ ▷ $\boldsymbol{\alpha}_i \triangleq \{-d_{\mathcal{S}}(\boldsymbol{\mathcal{G}}_i, \overline{\boldsymbol{\mathcal{G}}}_j)\}_{j\in[d]}$

while not converged **do**

  $\widehat{\boldsymbol{\mathcal{G}}} \leftarrow \text{ComputeReconstructedSpectralDecomposition}(\boldsymbol{\alpha}, \overline{\boldsymbol{\mathcal{G}}})$ ▷ *Equation* (5)

  $\boldsymbol{\mathcal{L}} \leftarrow \text{ComputeDataFittingLoss}(\widehat{\boldsymbol{\mathcal{G}}}, \boldsymbol{\mathcal{G}})$ ▷ *Equation* (1)

  $\mathbf{R}, \boldsymbol{\Lambda}, \mathbf{L} \leftarrow \text{UpdateCoefficientLogit}(\boldsymbol{\mathcal{L}}, \boldsymbol{\alpha})$

return $\boldsymbol{\alpha}$

Algorithm 2 Learn DOODL dictionary

Require: $\boldsymbol{\mathcal{G}} \triangleq \{\boldsymbol{\mathcal{G}}_i\}_{i\in[N]} \in \mathcal{M}^N, \overline{\boldsymbol{\mathcal{G}}} \triangleq \{\overline{\boldsymbol{\mathcal{G}}}_i\}_{i\in[d]} \in \mathcal{M}^d$ ▷ *Operator / Dictionary*

while not converged **do**

  $\boldsymbol{\alpha} \leftarrow \text{EstimateOperatorCoefficient}(\boldsymbol{\mathcal{G}}, \overline{\boldsymbol{\mathcal{G}}})$ ▷ *Algorithm 1*

  $\widehat{\boldsymbol{\mathcal{G}}} \leftarrow \text{ComputeReconstructedSpectralDecomposition}(\boldsymbol{\alpha}, \overline{\boldsymbol{\mathcal{G}}})$ ▷ *Equation* (5)

  $\boldsymbol{\mathcal{L}} \leftarrow \text{ComputeDataFittingLoss}(\widehat{\boldsymbol{\mathcal{G}}}, \boldsymbol{\mathcal{G}})$ ▷ *Equation* (1)

  $\boldsymbol{\xi} \leftarrow \text{ComputeAmbientSpaceDictionaryGradient}(\boldsymbol{\mathcal{L}})$ ▷ $\widehat{\boldsymbol{\mathcal{G}}}$ *is fixed*

  $\boldsymbol{\xi} \leftarrow \text{ComputeRiemannianGradient}(\boldsymbol{\xi}, \overline{\boldsymbol{\mathcal{G}}})$ ▷ *Proposition 5*

  $\overline{\boldsymbol{\mathcal{G}}} \leftarrow \text{UpdateDictionary}(\boldsymbol{\xi}, \overline{\boldsymbol{\mathcal{G}}})$ ▷ *Perform a retraction, see Proposition 1*

return $\overline{\boldsymbol{\mathcal{G}}}$

D Dictionary Learning on Langevin dynamics

D.1 Langevin dynamics

Langevin dynamics provides a fundamental mathematical framework for modeling stochastic dynamical systems evolving under the combined influence of deterministic forces and thermal or environmental noise. It underpins a wide range of scientific domains, including molecular dynamics, statistical physics, chemistry, materials discovery, and biological conformational analysis, where the key quantities of interest are inherently dynamical and spectral: metastable states, transition pathways, relaxation time scales, and invariant distributions [55, 56]. In these settings, obtaining reliable estimates of the underlying evolution operators is notoriously difficult because trajectories are often short, high-dimensional, partially observed, and collected far from equilibrium [6, 64].

D.2 Experimental settings

The whole experiment is ran on a single GPU (NVIDIA RTX A6000).

Trajectory simulation We consider a one-dimensional damped Langevin dynamics governed by $\mathrm{d}X_t = \nabla U_w(X_t) + \sqrt{2\sigma}\mathrm{d}B_t$ with $\{B_t\}_{t>0}$ a Brownian motion, $\sigma > 0$ its temperature, and a two well potential $U_w(x) \triangleq w^{-4}(x^2 - w^2)^2$ where $w > 0$ controls the distance between the two wells and their respective width, see fig. 1-left. We uniformly sampled 256/256 train/test potential parameters $w \in [0.5, 1.2]$ and generates trajectories of 40k samples at 100Hz according to their Langevin dynamics and with the Euler–Maruyama method.

Operator estimation. The operators are estimated via Reduced Rank Regression (RRR) [34] with a tikhonov regularization of 1e-6. Operators rank is fixed to 3 and the observable space approximates the RBF kernel via 400 Random Fourier Features [4] on sliding windows of 50 samples.

Dictionary learning. On the training set, we learn DOODL dictionaries eq. (4) with 2 to 5 atoms, using the SGOT divergence with $\eta = 0.25$ and the log-Martin metric for the spectral projector term. The training of each dictionary last for 3 epochs with batch sizes of 32. Dictionary is learned following the procedure described in Appendix section C.2 with the default settings.

D.3 Additional results.

We provide additional results for the estimation of operators from short trajectories in Figure 7 and Figure 8 by including DOODL dictionaries with 3 and 4 atoms. We observe the DOODL performs better than other estimators for trajectories up to 10k samples with lower variances for all potential widths. As well by increasing the number of atoms operator estimation with DOODL improves in accuracy.

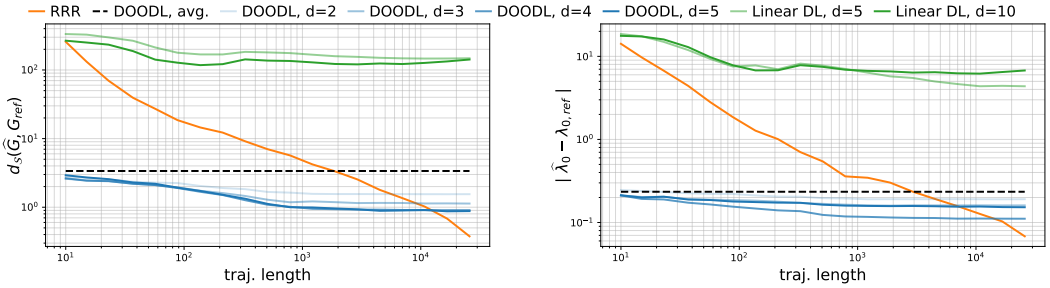

Figure 7: Comparison of operator estimator on truncated trajectories, including individual RRR, linear DL and DOODL. Error to ground truth in SGOT divergence between operators (left) and absolute error between first eigenvalues (right).

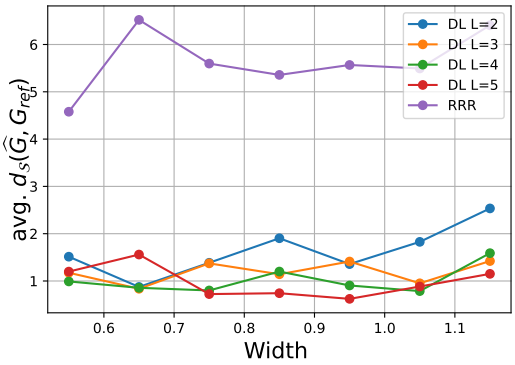

Figure 8: Comparison of operator estimation per potential width with SGOT error on trajectories of length 1.2k. Including individual RRR and DOODL.

E Plasma Experiments

This section details the plasma pipeline used in the Tokam2D experiments. Starting from high-dimensional plasma trajectories, we construct finite-rank generator representations and then test whether their operator geometry retains the physical parameters of the dynamicals systems. The pipeline is

$$\mathbf{x}_t \longmapsto \psi_{\mathrm{P}}(\mathbf{x}_t) \longmapsto \mathbf{z}_t \longmapsto \widehat{G} \longmapsto \widehat{\boldsymbol{\alpha}}.$$

Here $\mathbf{x}_t$ is a four-channel plasma snapshot, $\psi_{\mathrm{P}}(\mathbf{x}_t)$ is the frozen POSEIDON descriptor, $\mathbf{z}_t$ is the learned temporal coordinate, $\widehat{G} \in \mathcal{S}_r(\mathcal{H})$ is the finite-rank generator representation, and $\widehat{\boldsymbol{\alpha}} \in \Delta_K$ is the DOODL barycentric coordinate.

The parameters (g, κ) are never used during POSEIDON feature extraction, temporal representation learning, generator-resolvent estimation, or dictionary learning (All the pipeline is Unsupervised). They are used only afterward as diagnostics. Thus, recovering (g, κ) from the learned operator coordinates indicates that the unsupervised operator geometry captures physical structure in the Tokam2D regime family.

Note that the whole experiment is ran on a single GPU (NVIDIA RTX A6000).

E.1 Plasma details.

The plasma benchmark is based on a reduced two-field model for edge turbulence with density $n(x, y, t)$ and electrostatic potential $\phi(x, y, t)$ [18, 19]. The vorticity is $W = \Delta_\perp \phi$, with $\Delta_\perp = \partial_{xx} + \partial_{yy}$. The potential also defines the advecting velocity through its spatial derivatives $(-\partial_y \phi, \partial_x \phi)$, which are used as input channels in the observation map. The nonlinear advection is written with the Poisson bracket

$$[\phi, f] = \partial_x \phi \, \partial_y f - \partial_y \phi \, \partial_x f.$$

We use the gradient-driven configuration, where a prescribed background density gradient replaces a localized source. This gives the scalar control parameter $\kappa = 1/L_n$. The dynamics are

$$\begin{cases} \partial_t n + \kappa \partial_y \phi + [\phi, n] - D_n \Delta_\perp n = -\sigma_n n + \sigma_{n\phi} \phi, \\ \partial_t W + g \partial_y n + [\phi, W] - D_\phi \Delta_\perp W = -\sigma_{\phi n} n + \sigma_\phi \phi, \qquad W = \Delta_\perp \phi. \end{cases}$$

The parameter κ controls the density-gradient drive in the density equation, while g controls the interchange coupling in the vorticity equation through $g\partial_y n$. Thus, κ acts directly on the transported density field, whereas g acts through the density–vorticity coupling and modifies the potential-driven flow. In the linear interchange analysis, the drive depends multiplicatively on g and the background gradient, which explains why the strongest changes are observed when both parameters are large [18]. All diffusion and loss coefficients are fixed.

We generate the data with Tokam2D [21], a reduced 2D fluid solver in the plane transverse to the magnetic field. We uniformly sample $M = 400$ dynamics with $(g, \kappa) \in [0, 0.5] \times [1, 3.5]$. Each system is simulated on a 128×128 grid for $T = 4093$ time steps with $\Delta t = 0.05$, and we use an $80/20$ train/test split over simulations. Each regime m is represented by $X^{(m)} = (x_1^{(m)}, \ldots, x_T^{(m)})$, with

$$x_t^{(m)} = \left(n_t^{(m)}, -\partial_y \phi_t^{(m)}, \partial_x \phi_t^{(m)}, \phi_t^{(m)}\right) \in \mathbb{R}^{4 \times 128 \times 128}.$$

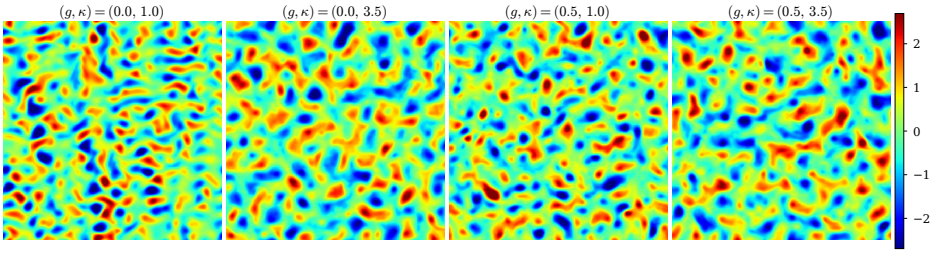

Figure 9: electrostatic potential $\phi(x, y, t)$ fluctuations at the four corners of the (g, κ) grid. In contrast to density snapshots, which mainly show the transported scalar field, the potential highlights the organization of the advecting flow. Increasing g makes the potential structures more deformed and vortex-like, consistently with its role in the vorticity equation.

Figure 9 complements the density snapshots by showing the potential-driven flow organization. The panels are standardized independently, so they compare morphology rather than amplitude. Increasing κ changes the spatial scale, whereas increasing g makes the structures more distorted and vortex-like, especially in the high-κ regime. This suggests that part of the information on g is carried by the potential/vorticity dynamics rather than by density fluctuations alone.

E.2 POSEIDON as a PDE foundation model encoder

The first stage maps each plasma state to a finite-dimensional PDE-aware descriptor. We use POSEIDON-B [24] as a frozen feature extractor. POSEIDON is a PDE foundation model pretrained to approximate solution operators: given a state $\mathbf{u}(t_i)$ and a lead time τ, its native task is

$$\mathcal{S}_\theta(\tau, \mathbf{u}(t_i)) \simeq \mathbf{u}(t_i + \tau).$$

Its scOT backbone uses patch embeddings, multiscale shifted-window attention, and lead-time conditioning. The POSEIDON paper also reports strong transfer to out-of-distribution PDEs and physical processes not seen during pretraining. This is the setting we rely on here: POSEIDON is not

used to forecast Tokam2D trajectories. The recovery head is discarded, all POSEIDON parameters are frozen, and only the internal encoder representation is retained.

The choice of channels is guided by the variables and differential operators appearing in the equations. Define

$$\nabla^\perp \phi := (-\partial_y \phi, \partial_x \phi), \qquad \mathbf{V}_E := \nabla^\perp \phi, \qquad \nabla \cdot \mathbf{V}_E = 0.$$

Then

$$[\phi, f] = \partial_x \phi \, \partial_y f - \partial_y \phi \, \partial_x f = \mathbf{V}_E \cdot \nabla f,$$

and the Tokam2D system can be written as

$$\begin{cases} \partial_t n + \mathbf{V}_E \cdot \nabla n + \kappa \partial_y \phi - D_n \Delta_\perp n = -\sigma_n n + \sigma_{n\phi} \phi, \\ \partial_t W + \mathbf{V}_E \cdot \nabla W + g \partial_y n - D_\phi \Delta_\perp W = -\sigma_{\phi n} n + \sigma_\phi \phi, \qquad W = \Delta_\perp \phi. \end{cases}$$

Thus Tokam2D has the same basic transport form as the incompressible equations used in POSEIDON pretraining: a divergence-free vector field advects scalar or vorticity-like quantities, with diffusion and coupling terms. In two-dimensional incompressible Navier–Stokes, writing $\mathbf{u} = (u, v)$, one has

$$\partial_t \mathbf{u} + (\mathbf{u} \cdot \nabla)\mathbf{u} + \nabla p - \nu \Delta \mathbf{u} = 0, \qquad \nabla \cdot \mathbf{u} = 0,$$

or equivalently in vorticity form

$$\partial_t \omega + \mathbf{u} \cdot \nabla \omega - \nu \Delta \omega = 0, \qquad \mathbf{u} = \nabla^\perp \psi, \qquad \omega = \Delta \psi.$$

The corresponding structures are

$$\mathbf{V}_E = \nabla^\perp \phi \quad \leftrightarrow \quad \mathbf{u} = \nabla^\perp \psi, \qquad W = \Delta_\perp \phi \quad \leftrightarrow \quad \omega = \Delta \psi.$$

Moreover, POSEIDON is also pretrained on compressible Euler variables, which include a density ρ, a velocity field (u, v), and a pressure-like scalar p. We therefore feed POSEIDON the four-channel Tokam2D state

$$\mathbf{x}_t^{(m)} = \big(n_t^{(m)}, V_{E,x,t}^{(m)}, V_{E,y,t}^{(m)}, \phi_t^{(m)}\big) = \big(n_t^{(m)}, -\partial_y \phi_t^{(m)}, \partial_x \phi_t^{(m)}, \phi_t^{(m)}\big) \in \mathbb{R}^{4 \times 128 \times 128},$$

with the channel correspondence

$$n_t \leftrightarrow \rho, \qquad (V_{E,x,t}, V_{E,y,t}) \leftrightarrow (u, v), \qquad \phi_t \leftrightarrow p.$$

This gives POSEIDON fields with the same mathematical roles as in its pretraining interface: a density-like scalar, a vector transport field, and a scalar potential or pressure-like channel.

Before applying the frozen encoder, each channel is standardized independently for each snapshot. The normalized four-channel field is passed through the frozen POSEIDON patch embedding and scOT encoder. If

$$\mathbf{H}_t^{(m)} = \mathrm{Enc}_{\mathrm{P}}(\widetilde{\mathbf{x}}_t^{(m)}) \in \mathbb{R}^{N_{\mathrm{tok}} \times 768}$$

denotes the last hidden token matrix, we define the frozen POSEIDON descriptor by token averaging,

$$\psi_{\mathrm{P}}(\mathbf{x}_t^{(m)}) = \frac{1}{N_{\mathrm{tok}}} \sum_{\ell=1}^{N_{\mathrm{tok}}} \mathbf{H}_{t,\ell}^{(m)} \in \mathbb{R}^{768}.$$

For each regime, this gives the descriptor trajectory

$$\mathbf{\Psi}_{\mathrm{P}}^{(m)} = \big(\psi_{\mathrm{P}}(\mathbf{x}_1^{(m)}), \ldots, \psi_{\mathrm{P}}(\mathbf{x}_T^{(m)})\big) \in \mathbb{R}^{T \times 768}.$$

These descriptors are precomputed once for every simulation. The operators used by DOODL are not estimated directly from $\mathbf{\Psi}_{\mathrm{P}}^{(m)}$, but from the learned temporal coordinates described next.

E.3 Observable space learning for single dynamical system

The frozen POSEIDON descriptors encode spatial morphology, but they are not optimized for spectral operator estimation. We therefore learn temporal coordinates using the contrastive loss from Neural Conditional Probability (NCP) [36]. We use the same loss as NCP, but not for conditional probability estimation: here, it is used only to extract a latent space adapted to the plasma dynamics.

For each regime m, we sample temporal windows $\{(t_i, t_i+1)\}_{i=1}^n$ and apply two shared MLP heads

$$u_\theta : \mathbb{R}^{768} \to \mathbb{R}^{d_z}, \qquad v_\theta : \mathbb{R}^{768} \to \mathbb{R}^{d_z}, \qquad d_z = 64.$$

The current and future coordinates are

$$\mathbf{Z}_-^{(m)} = \left[u_\theta(\psi_{\mathrm{P}}(\mathbf{x}_{t_1}^{(m)})), \dots, u_\theta(\psi_{\mathrm{P}}(\mathbf{x}_{t_n}^{(m)}))\right] \in \mathbb{R}^{d_z \times n},$$

$$\mathbf{Z}_+^{(m)} = \left[v_\theta(\psi_{\mathrm{P}}(\mathbf{x}_{t_1+1}^{(m)})), \dots, v_\theta(\psi_{\mathrm{P}}(\mathbf{x}_{t_n+1}^{(m)}))\right] \in \mathbb{R}^{d_z \times n}.$$

We define the score matrix

$$\mathbf{S}^{(m)} = \left(\mathbf{Z}_-^{(m)}\right)^\top \mathbf{Z}_+^{(m)}, \qquad S_{ij}^{(m)} = \left\langle u_\theta(\psi_{\mathrm{P}}(\mathbf{x}_{t_i}^{(m)})), v_\theta(\psi_{\mathrm{P}}(\mathbf{x}_{t_j+1}^{(m)}))\right\rangle.$$

The NCP contrastive loss for regime m is

$$\ell_m(\theta) = \frac{1}{n(n-1)} \sum_{i \neq j} \left(S_{ij}^{(m)}\right)^2 - \frac{2}{n} \sum_{i=1}^n S_{ii}^{(m)} + \gamma \left[r(\mathbf{Z}_-^{(m)}) + r(\mathbf{Z}_+^{(m)})\right]. \tag{100}$$

The diagonal term aligns true one-step pairs, while the off-diagonal term penalizes mismatched temporal pairs.

The regularizer prevents collapse and keeps the coordinates well conditioned. For $\mathbf{Z} \in \mathbb{R}^{d_z \times n}$, define

$$\bar{\mathbf{z}} = \frac{1}{n}\mathbf{Z}\mathbf{1}, \qquad \widehat{\mathbf{\Sigma}}_{\mathbf{Z}} = \frac{1}{n}(\mathbf{Z} - \bar{\mathbf{z}}\mathbf{1}^\top)(\mathbf{Z} - \bar{\mathbf{z}}\mathbf{1}^\top)^\top.$$

We use

$$r(\mathbf{Z}) = \frac{1}{d_z}\|\widehat{\mathbf{\Sigma}}_{\mathbf{Z}} - \mathbf{I}\|_F^2 + 2\|\bar{\mathbf{z}}\|_2^2. \tag{101}$$

This centers and approximately whitens the learned coordinates over each temporal window.

E.4 Observable space learning on multiple dynamical systems with GradNorm multitask strategy.

Multitask balancing across plasma regimes. Each Tokam2D regime m corresponds to a parameter pair $(g^{(m)}, \kappa^{(m)})$ and is treated as one task. Let $\mathcal{M}_{\mathrm{tr}}$ be the set of training regimes and $\ell_m(\theta)$ the temporal representation loss of regime m. A uniform objective, $|\mathcal{M}_{\mathrm{tr}}|^{-1}\sum_{m\in\mathcal{M}_{\mathrm{tr}}} \ell_m(\theta)$, can be poorly balanced because regimes have different temporal complexity. The parameter κ controls the density-gradient drive, while g acts through the interchange coupling between density and vorticity; regimes where both effects are large can exhibit stronger fluctuations and richer temporal spectra.

We therefore use a GradNorm-weighted multitask objective [11]. For a mini-batch of regimes $\mathcal{B} \subset \mathcal{M}_{\mathrm{tr}}$,

$$\mathcal{L}_{\mathrm{MTL}}(\theta, \omega) = \sum_{m\in\mathcal{B}} \omega_m \ell_m(\theta), \qquad \omega_m = |\mathcal{M}_{\mathrm{tr}}| \frac{\exp(\beta_m)}{\sum_{j\in\mathcal{M}_{\mathrm{tr}}} \exp(\beta_j)}, \tag{102}$$

so that the average task weight over training regimes is one. The logits β_m are learned with GradNorm, while θ denotes the shared temporal-head parameters.

Let $\ell_m(0)$ be the initial loss of regime m. At iteration t, define the relative training rate

$$r_m(t) = \frac{\ell_m(t)/\ell_m(0)}{|\mathcal{B}|^{-1}\sum_{j\in\mathcal{B}} \ell_j(t)/\ell_j(0)}.$$

Thus, $r_m(t) > 1$ means that regime m is learning more slowly than the mini-batch average. Let W_{ref} be the last shared layer of the temporal heads. GradNorm measures

$$G_m(t) = \left\|\nabla_{W_{\mathrm{ref}}}\left(\omega_m \ell_m(t)\right)\right\|_2, \qquad \widetilde{G}_m(t) = \overline{G}(t) r_m(t)^{\alpha_{\mathrm{GN}}},$$

where

$$\overline{G}(t) = |\mathcal{B}|^{-1} \sum_{j\in\mathcal{B}} G_j(t), \qquad \alpha_{\mathrm{GN}} = 0.12.$$

The task weights are updated by minimizing

$$\mathcal{L}_{\mathrm{GN}} = \sum_{m \in \mathcal{B}} \left| G_m(t) - \widetilde{G}_m(t) \right|, \tag{103}$$

and the shared representation parameters are updated with $\mathcal{L}_{\mathrm{MTL}}$.

This balancing does not force all regimes to be equally easy. It prevents regimes with larger loss scale or stronger turbulent activity from dominating the shared representation. In practice, higher weights on high-drive or strongly coupled regimes are consistent with their richer temporal spectra and help learn coordinates that remain usefu.

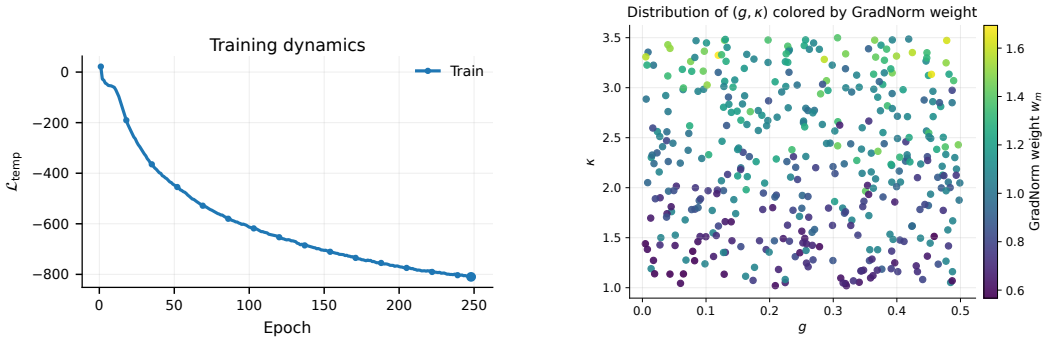

Figure 10: Temporal representation learning diagnostics. Left: training objective in (100). Right: final GradNorm weights over the (g, κ) map. Larger weights indicate regimes that require more optimization effort in the shared temporal representation.

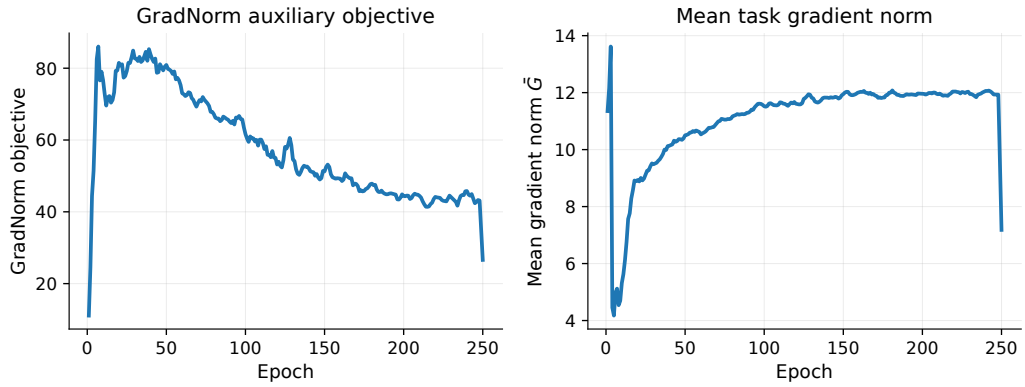

Figure 11: GradNorm optimization diagnostics. Left: auxiliary GradNorm objective. Right: gradient norm of the main multitask objective. The curves show that the task weights stabilize during training.

The final GradNorm weights in Figure E.4 indicate where the shared temporal representation has more difficulty fitting all regimes simultaneously. The largest weights are concentrated at high κ, matching the density snapshots where larger κ produces stronger, more organized fluctuations. Increasing g alone has a weaker effect on the weights, consistent with its more indirect role through the density–vorticity coupling. Its influence becomes more visible at high κ, where the potential snapshots show more distorted flow structures. Overall, the weighting diagnostic suggests that the hardest regimes are those with stronger turbulent activity and richer temporal content.

Reduced-rank validation of the temporal representation. We use reduced-rank regression (RRR) only as a validation diagnostic for the temporal representation; it is not the spectral estimator used by DOODL. The goal is to check whether the learned coordinates admit a predictable finite-rank one-step evolution.

For each regime m, we form one-step pairs in the learned coordinate space,

$$\mathbf{z}_i^{-,(m)} = u_\theta\Big(\psi_{\mathrm{P}}(\mathbf{x}_{t_i}^{(m)})\Big), \qquad \mathbf{z}_i^{+,(m)} = u_\theta\Big(\psi_{\mathrm{P}}(\mathbf{x}_{t_i+1}^{(m)})\Big).$$

For a candidate rank r, we fit a ridge-regularized reduced-rank predictor

$$\widehat{A}_{m,r} \in \arg\min_{\mathrm{rank}(A)\le r} \frac{1}{|\mathcal{I}_{\mathrm{tr}}^{(m)}|} \sum_{i\in\mathcal{I}_{\mathrm{tr}}^{(m)}} \left\| A\mathbf{z}_i^{-,(m)} - \mathbf{z}_i^{+,(m)} \right\|_2^2 + \lambda\|A\|_F^2. \tag{104}$$

We first select the diagnostic rank by evaluating the held-out one-step error

$$\mathrm{Err}_{\mathrm{RRR}}(r) = \frac{1}{|\mathcal{M}_{\mathrm{val}}|} \sum_{m\in\mathcal{M}_{\mathrm{val}}} \frac{1}{|\mathcal{I}_{\mathrm{val}}^{(m)}|} \sum_{i\in\mathcal{I}_{\mathrm{val}}^{(m)}} \left\| \widehat{A}_{m,r}\mathbf{z}_i^{-,(m)} - \mathbf{z}_i^{+,(m)} \right\|_2^2.$$

Figure 12 shows that the error decreases at small ranks and then saturates. We therefore use rank $r_{\mathrm{eval}} = 40$ for checkpoint selection.

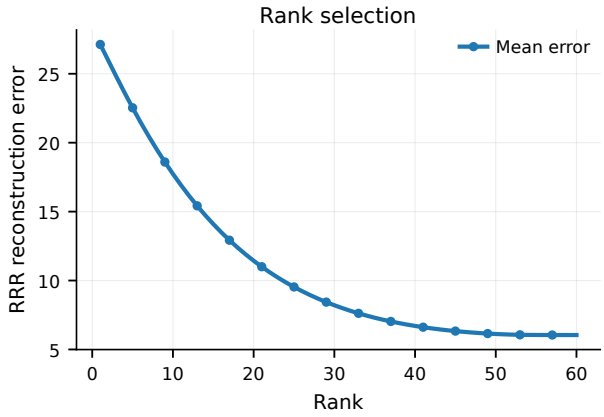

Figure 12: Diagnostic rank selection for reduced-rank validation. The held-out one-step error decreases at small ranks and then saturates, indicating that a moderate-rank predictor captures most of the predictable temporal structure in the learned coordinates.

Given $r_{\mathrm{eval}} = 40$, we monitor the one-step coefficient of determination

$$R^2_{\mathrm{RRR}} = 1 - \frac{\sum_m \sum_{i\in\mathcal{I}^{(m)}} \left\| \widehat{A}_{m,r_{\mathrm{eval}}}\mathbf{z}_i^{-,(m)} - \mathbf{z}_i^{+,(m)} \right\|_2^2}{\sum_m \sum_{i\in\mathcal{I}^{(m)}} \left\| \mathbf{z}_i^{+,(m)} - \bar{\mathbf{z}}^{+,(m)} \right\|_2^2}, \qquad \bar{\mathbf{z}}^{+,(m)} = \frac{1}{|\mathcal{I}^{(m)}|} \sum_{i\in\mathcal{I}^{(m)}} \mathbf{z}_i^{+,(m)}.$$

This score measures how much one-step variation is explained by a finite-rank linear predictor, relative to a constant predictor. We compute it on both training and validation regimes and select the checkpoint maximizing validation R^2_{RRR}.

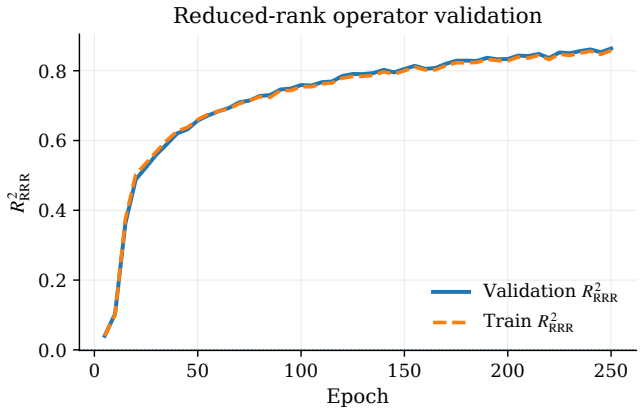

Figure 13: Checkpoint selection with reduced-rank validation. Train and validation R^2_{RRR} measure finite-rank one-step predictability in the learned coordinates. The selected checkpoint maximizes validation R^2_{RRR}, while the train/validation comparison checks that the representation improves without relying on physical labels.

E.5 Generator-resolvent estimation

For each regime m, the temporal representation step produces the latent trajectory

$$\mathbf{z}^{(m)}_{1:T} = \big(\mathbf{z}^{(m)}_1, \dots, \mathbf{z}^{(m)}_T\big), \qquad \mathbf{z}^{(m)}_t = u_\theta\Big(\psi_{\mathrm{P}}(\mathbf{x}^{(m)}_t)\Big) \in \mathbb{R}^{d_z}.$$

From this trajectory, we estimate a finite-rank spectral generator representation

$$\widehat{G}_m = (\widehat{\Lambda}_m, \widehat{L}_m, \widehat{R}_m)$$

using a generator-resolvent (GR) estimator [35]. The GR estimator is well suited to plasma trajectories because it aggregates time-lagged information over multiple lags. This helps capture dynamics occurring at different time scales, including oscillatory drift/interchange activity and slower transport structures, while producing a stable spectral representation of the generator.

The estimated eigenvalues $\widehat{\lambda}_i = \widehat{\sigma}_i + \mathrm{i}\widehat{\omega}_i$ encode growth or decay through $\widehat{\sigma}_i$ and oscillatory content through $\widehat{\omega}_i$, while the spectral factors $(\widehat{L}_m, \widehat{R}_m)$ encode the associated modes. The resulting $\widehat{G}_m$ is the operator representation passed to DOODL. Details of the generator-resolvent construction are given in Section A.1.

Choice of shift, lag, and rank. For the plasma experiments, each learned latent trajectory $\mathbf{z}^{(m)}_{1:T} = (\mathbf{z}^{(m)}_1, \dots, \mathbf{z}^{(m)}_T)$ is mapped to a finite-rank spectral generator representation

$$\widehat{G}_m = (\widehat{\Lambda}_m, \widehat{L}_m, \widehat{R}_m),$$

where $\widehat{\Lambda}_m$ contains the recovered generator eigenvalues and $\widehat{L}_m, \widehat{R}_m$ are the associated left and right spectral factors. We use a generator-resolvent (GR) filter with

$$a = 8, \qquad K = 600, \qquad r_{\mathrm{GR}} = 40,$$

Here a is the GR shift, K is the maximum lag used by the time-lagged filter, and r_{GR} is the number of retained spectral components. The eigenvalues $\widehat{\nu}_i$ of the GR-filtered operator are mapped back to generator eigenvalues by

$$\widehat{\lambda}_i = a\left(1 - \widehat{\nu}_i^{-1}\right).$$

All spectra shown below are therefore spectra of the recovered generator, not spectra of the resolvent.

The shift a controls the resolution–stability tradeoff of the recovered generator spectrum. Small shifts reveal finer spectral variation but are more sensitive to numerical noise, whereas large shifts smooth the estimate but compress the recovered eigenvalues. The value $a = 8$ gives a stable compromise for the downstream SGOT geometry. The maximum lag controls the temporal support of the GR

estimate: with $\Delta t = 0.05$, $K = 600$ corresponds to a time window of length 30. This provides enough temporal context to capture slow transport and oscillatory content while remaining stable across regimes. Finally, $r_{\mathrm{GR}} = 40$ retains a rich spectral description, including oscillatory and subspace information useful for DOODL, without making the representation dominated by weak numerical modes.

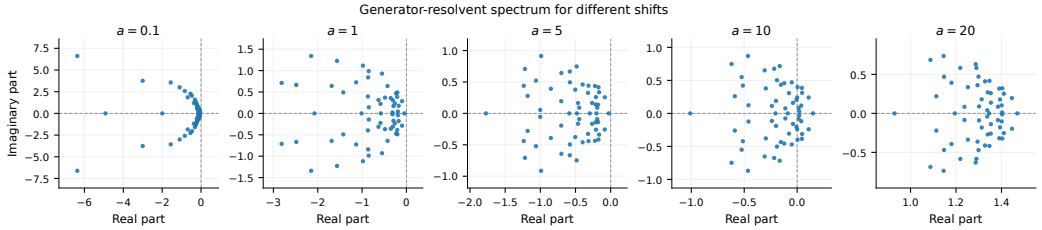

Figure 14: Recovered generator spectra for different GR shifts a, with fixed maximum lag. The plotted points are generator eigenvalues obtained after inversion of the GR filter. Small shifts reveal finer spectral variation, while larger shifts produce smoother but more compressed spectra. We use $a = 8$ in the plasma pipeline.

Figure 14 shows that the recovered generator spectra remain mostly in the stable half-plane and retain nonzero imaginary components. This is consistent with the operator structure expected for statistically stable plasma regimes: negative real parts encode decay or relaxation, while imaginary parts encode oscillatory drift/interchange content. Since the latent dynamics are real-valued, complex eigenvalues appear in conjugate pairs up to numerical error.

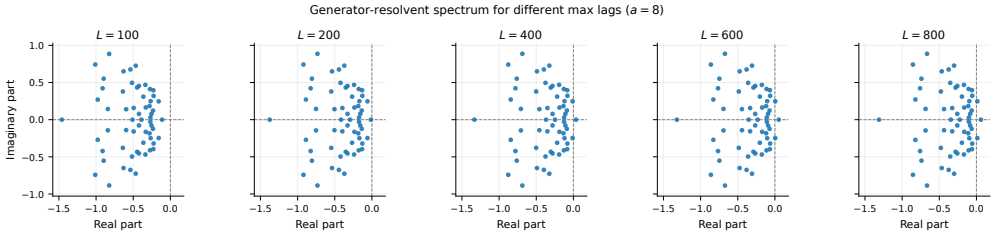

Figure 15: Recovered generator spectra for different maximum lags K, with fixed GR shift $a = 8$. The plotted points are generator eigenvalues obtained after inversion of the GR filter. The dominant spectral organization is stable across the tested lags, supporting the use of $K = 600$.

Figure 15 shows that the main spectral structure is stable across the tested lag values. We therefore use $K = 600$, which provides enough temporal context for slow modes without visibly destabilizing the recovered spectrum.

Generator-resolvent eigenfunction prediction score. We use an eigenfunction prediction score as a dynamic consistency diagnostic. Let $\widehat{\lambda}_i$ be the recovered generator eigenvalues and $\widehat{\psi}_i(t)$ the corresponding right eigenfunction evaluations along the trajectory. A consistent generator estimate should satisfy

$$\widehat{\psi}_i(t+1) \simeq e^{\Delta t\, \widehat{\lambda}_i} \widehat{\psi}_i(t).$$

We report

$$R^2_{\mathrm{GR}} = 1 - \frac{\sum_{t=1}^{T-1} \sum_{i=1}^{r_{\mathrm{GR}}} \left| \widehat{\psi}_i(t+1) - e^{\Delta t\, \widehat{\lambda}_i} \widehat{\psi}_i(t) \right|^2}{\sum_{t=1}^{T-1} \sum_{i=1}^{r_{\mathrm{GR}}} \left| \widehat{\psi}_i(t+1) - \bar{\psi}_i \right|^2}, \qquad \bar{\psi}_i = \frac{1}{T-1} \sum_{t=1}^{T-1} \widehat{\psi}_i(t+1).$$

This score measures whether the recovered spectral coordinates evolve according to the estimated generator eigenvalues. . In the selected configuration, we obtain $R^2_{\mathrm{GR}} = 0.97$,

E.6 Spectral coordinates and physical parameters

We finally inspect whether the recovered generator spectra retain the physical parameters of the Tokam2D dynamical systems from unsupervised learning.

For each regime $m \in [M]$, the generator-resolvent estimator returns generator eigenvalues $\widehat{\lambda}_i^{(m)} = \widehat{\sigma}_i^{(m)} + \mathrm{i}\widehat{\omega}_i^{(m)}$, $i = 1, \dots, r_{\mathrm{GR}}$. We encode the recovered spectrum by

$$\mathbf{s}^{(m)} = \left(\widehat{\sigma}_1^{(m)}, \dots, \widehat{\sigma}_{r_{\mathrm{GR}}}^{(m)}, |\widehat{\omega}_1^{(m)}|, \dots, |\widehat{\omega}_{r_{\mathrm{GR}}}^{(m)}|\right)^\top \in \mathbb{R}^{2r_{\mathrm{GR}}}.$$

The real parts encode decay or growth rates, while the absolute imaginary parts encode frequency scales. We use absolute values because non-real eigenvalues appear in complex conjugate pairs. Stacking all regimes gives a matrix $\mathbf{S} \in \mathbb{R}^{M \times 2r_{\mathrm{GR}}}$, whose columns are standardized across regimes. The parameter vectors $\mathbf{g}, \boldsymbol{\kappa} \in \mathbb{R}^M$ are standardized as well.

Marginal spectral associations. For $\mathbf{y} \in \{\mathbf{g}, \boldsymbol{\kappa}\}$, we compute the marginal association between the j-th spectral coordinate and $\mathbf{y}$ as

$$\rho_{y,j} = \frac{\langle \mathbf{S}_{:,j}, \mathbf{y} \rangle}{\|\mathbf{S}_{:,j}\|_2 \|\mathbf{y}\|_2}, \qquad j = 1, \dots, 2r_{\mathrm{GR}}.$$

Thus, $\rho_{y,j}$ is the empirical correlation between a recovered generator coordinate and the physical parameter.

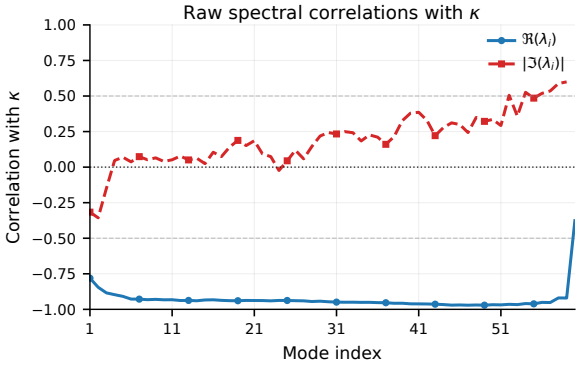

Figure 16: Marginal correlations between recovered generator coordinates and the density-gradient parameter κ. Large coefficients identify decay-rate or frequency coordinates that vary systematically with the imposed density-gradient drive.

Figure 16 shows a strong spectral signature of κ. This is consistent with the model: κ directly controls the density-gradient drive, and therefore changes the dominant activity level, spatial scale, and temporal organization of the plasma dynamics.

Residual association with g after regressing out κ. The interchange parameter g has a weaker marginal spectral signature than κ. This is consistent with the Tokam2D equations: κ enters as the density-gradient drive, whereas g acts through the coupling between density perturbations and vorticity. Its effect can therefore be partially masked by the dominant variation induced by κ.

To isolate the part of the recovered spectrum associated with g beyond this dominant κ-direction, we remove the linear effect of κ from both g and each spectral coordinate. Let $\mathbf{S} \in \mathbb{R}^{M \times 2r_{\mathrm{GR}}}$ be the standardized matrix of spectral coordinates, where the j-th column $\mathbf{S}_{:,j}$ contains one coordinate across the M regimes. We also denote by $\boldsymbol{\kappa}, \mathbf{g} \in \mathbb{R}^M$ the standardized parameter vectors.

For each spectral coordinate j, we fit the linear regression

$$\mathbf{S}_{:,j} = \beta_{0,j}\mathbf{1} + \beta_{1,j}\boldsymbol{\kappa} + \mathbf{r}_j^{(\kappa)},$$

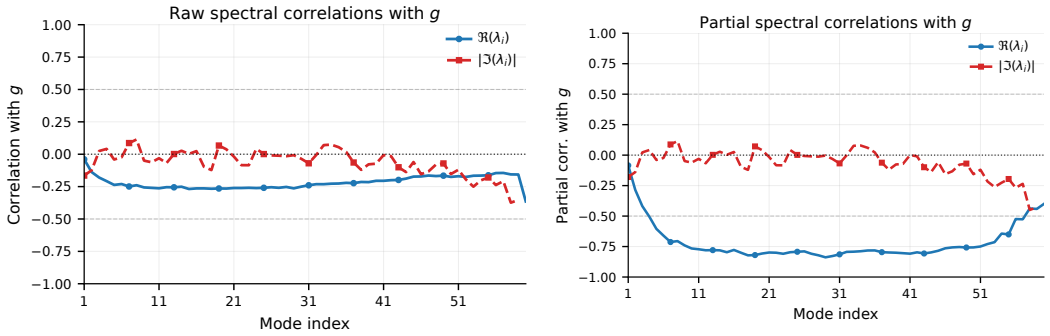

Figure 17: Spectral correlations with the interchange parameter g. Left: marginal correlations between g and the recovered generator coordinates. Right: partial correlations obtained after regressing out the linear effect of κ from both g and each spectral coordinate. The residual signal shows that the recovered generator spectrum retains information about g beyond the dominant κ-driven variation.

where $\mathbf{r}_j^{(\kappa)}$ is the residual part of the j-th generator coordinate after removing its linear dependence on κ. We similarly regress g on $(\mathbf{1}, \boldsymbol{\kappa})$,

$$\mathbf{g} = \gamma_0 \mathbf{1} + \gamma_1 \boldsymbol{\kappa} + \mathbf{r}_g^{(\kappa)}.$$

We then define

$$\rho_{g|\kappa,j} = \frac{\left\langle \mathbf{r}_j^{(\kappa)}, \mathbf{r}_g^{(\kappa)} \right\rangle}{\left\| \mathbf{r}_j^{(\kappa)} \right\|_2 \left\| \mathbf{r}_g^{(\kappa)} \right\|_2}, \qquad j = 1, \ldots, 2r_{\mathrm{GR}}.$$

This is the empirical partial correlation between the j-th recovered generator coordinate and g, after removing the linear contribution of κ.

Figure 17 shows that the marginal spectral correlations with g are weaker than those for κ, but that a clear residual signal remains after regressing out κ. This indicates that the recovered generator spectrum does not only encode the dominant density-gradient drive. It also retains information about the interchange parameter through more distributed spectral coordinates, consistent with the role of g in the density–vorticity coupling.

Representative spectral coordinates. We visualize representative coordinates by selecting those with the largest values of $|\rho_{\kappa,j}|$ and $|\rho_{g|\kappa,j}|$. The first group corresponds to generator eigenvalues with the most important correlation with κ, while the second group captures the strongest residual correlation with g.

Figure 18 makes the previous correlations explicit. Coordinates associated with κ organize regimes along the main density-gradient direction. Coordinates associated with g are less dominant, but remain structured once the leading κ effect is removed. Overall, the recovered generator coordinates are therefore not arbitrary numerical features: they retain interpretable information about the physical control parameters of the Tokam2D dynamical systems.

SGOT geometry before dictionary projection. We also inspect the operator geometry before fitting the DOODL dictionary. From the generator-resolvent representations $\{\widehat{G}_m\}_{m=1}^{M}$, we compute the pairwise SGOT distance matrix using the log-Martin projector distance and spectral weight $\eta = 0.9$, then embed this distance matrix with t-SNE.

Figure 19 shows that the SGOT geometry of the recovered generators is already structured by the scanned parameters. The organization is strongest for κ, consistent with its direct role as the density-gradient drive, while g appears more weakly and in a less monotone way. This supports the role of DOODL as a compression of an already informative operator geometry, rather than a supervised mechanism for creating parameter information.

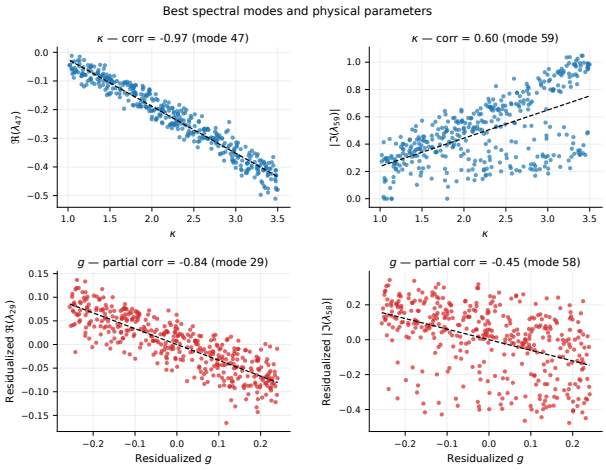

Figure 18: Representative recovered generator coordinates. Top: coordinates most associated with κ. Bottom: coordinates most associated with g after removing the effect of κ. The plots show that κ is encoded in dominant spectral directions, while g appears through weaker but structured residual variation.

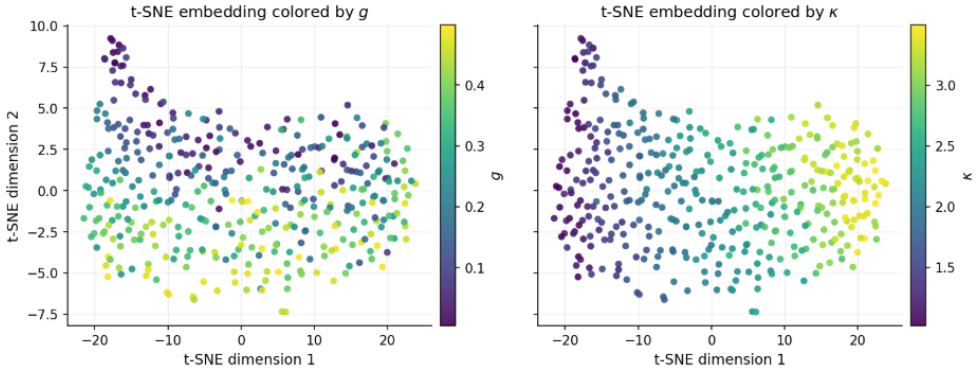

Figure 19: t-SNE embedding of the pairwise SGOT distance matrix between recovered generator operators, colored by the physical parameters g and κ. The embedding shows a clear organization along κ and a weaker but visible organization along g, indicating that physical information is already present in the operator geometry before dictionary learning.

E.7 Dictionary learning and parameter recovery

We apply DOODL to the full-trajectory generator representations $\widehat{G}_m = (\widehat{\Lambda}_m, \widehat{L}_m, \widehat{R}_m)$, $m = 1, \ldots, M$. The dictionary contains $d = 12$ atoms and is trained with the SGOT distance d_S, using the log-Martin projector distance and spectral weight $\eta = 0.9$.The training of each dictionary last for 6 epochs with batch sizes of 64. Dictionary is learned following the procedure described in section C.2 with the default settings. Each regime is represented by barycentric coordinates $\widehat{\boldsymbol{\alpha}}_m \in \Delta^{d-1}$ obtained by projecting $\widehat{G}_m$ onto the learned operator dictionary, i.e. by fitting the SGOT barycenter $B_{\widehat{\boldsymbol{\alpha}}_m}(D_1, \ldots, D_d)$ to $\widehat{G}_m$.

To test whether the unsupervised coordinates retain physical information, we fit two independent degree-2 polynomial ridge regressors on the training regimes, one from $\widehat{\boldsymbol{\alpha}}_m$ to $g^{(m)}$ and one from $\widehat{\boldsymbol{\alpha}}_m$ to $\kappa^{(m)}$.. As shown in Figure 5, regression from full-trajectory DOODL coordinates reaches $R^2_g = 0.811$ and $R^2_\kappa = 0.967$ on held-out regimes. The stronger recovery of κ is consistent with its dominant role as density-gradient drive, while the recovery of g shows that the dictionary also preserves information about the more indirect interchange dynamics.

E.8 Early operator and parameter recovery

We also evaluate short trajectories. For a trajectory length $\tau < T$, let $\widehat{\boldsymbol{\alpha}}_{m,\tau}$ denote the DOODL coordinates estimated from the short trajectory of regime m. These coordinates define the dictionary-constrained operator $\widehat{G}^{DOODL}_{m,\tau} = B_{\widehat{\boldsymbol{\alpha}}_{m,\tau}}(D_1, \ldots, D_d)$. We compare this estimate with a direct generator-resolvent estimate from the same short trajectory, using the SGOT distance to the full-trajectory operator $\widehat{G}_{m,T}$. As shown in Figure 6, the dictionary-constrained estimate is closer to $\widehat{G}_{m,T}$ in the short-data regime. This supports the interpretation of DOODL as an operator-level prior: early estimation is reduced to fitting coordinates in a learned spectral dictionary, rather than estimating an unconstrained generator from few observations.

For early parameter recovery, we train a separate diagnostic regressor. Full-trajectory recovery uses only the full-trajectory coordinates $\widehat{\boldsymbol{\alpha}}_m$. Early recovery uses data augmentation: the training set includes dictionary coordinates obtained from both full and short trajectories. For this early regressor, we use the pre-softmax dictionary logits $\widehat{\boldsymbol{\ell}}_{m,\tau} \in \mathbb{R}^d$ rather than the simplex-normalized coordinates, We again fit two independent degree-2 polynomial ridge regressors, one for g and one for κ. This augmentation is used only for the diagnostic parameter regressor; it does not change the learned operator dictionary.

Figure 6 shows that parameter recovery improves as longer trajectories become available. The recovery of κ rises quickly, consistent with its dominant spectral signature, while g improves more gradually, consistent with its more distributed residual signature in the generator spectrum. At the longest trajectory length, the early-recovery protocol reaches $R^2_g = 0.81$ and $R^2_\kappa = 0.97$. Figure 6 show that short trajectories already contain useful operator-level information, and that the learned DOODL dictionary makes this information recoverable in both SGOT geometry and physical parameter space.

E.9 Hyperparameters

Table 2: Main hyperparameters used in the Tokam2D plasma pipeline.

Parameter	Value
Number of regimes	400
Train/test split	$80/20$ over regimes
Grid size	128×128
Time step	$\Delta t = 0.05$
POSEIDON descriptor dimension	768
Temporal latent dimension	$d = 64$
Temporal heads u_θ, v_θ	two MLPs [512,256,64]
Task definition	one regime = one task
Task balancing	GradNorm
GradNorm exponent	$\alpha = 0.12$
Maximum epochs	250
Validation metric	R^2_{RRR}
RRR validation rank	$r_{\mathrm{eval}} = 40$
Selected validation score	$R^2_{\mathrm{RRR}} = 0.86$
Resolvent shift	$a = 8$
Maximum lag	$K = 600$
Generator resolvant rank	$r = 40$
GR validation score	$R^2_{\mathrm{GR}} = 0.97$
Regression diagnostic	degree-2 polynomial ridge
Full-trajectory regression scores	$R^2_g = 0.81,\ R^2_\kappa = 0.97$

F Statistical Guarantees

F.1 Setup and notation

Throughout we assume $(X_t)_{t\geq 1}$ is stationary with marginal π, so that $\mathrm{Var}_\pi(\ell_t(G))$ is time-independent for every fixed G. Let $\mathrm{z}_t := \phi(X_t) \in \mathcal{H}$ denote the encoded state under the feature map $\phi : \mathcal{X} \to \mathcal{H}$. Let $\bar{\boldsymbol{\mathcal{G}}} = (\bar{\boldsymbol{\mathcal{G}}}_j)_{j\in[d]} \in \mathcal{N}^d$ be a dictionary of d atoms, $\mathrm{B}(\boldsymbol{\alpha}; \bar{\boldsymbol{\mathcal{G}}}) \in \mathcal{N}$ the decoder (either B^{OT} or B^{p}), and

$$G_{\boldsymbol{\alpha}} := \mathrm{Op}[\mathrm{B}(\boldsymbol{\alpha}; \bar{\boldsymbol{\mathcal{G}}})] \in \mathcal{S}_r(\mathcal{H}), \qquad \mathcal{G}(\bar{\boldsymbol{\mathcal{G}}}) := \left\{G_{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \Delta^{d-1}\right\} \subset \mathcal{S}_r(\mathcal{H}). \tag{105}$$

The sample loss at operator $G \in \mathcal{S}_r(\mathcal{H})$ is $\ell_t(G) := \|\mathrm{z}_{t+1} - G^* \mathrm{z}_t\|_{\mathcal{H}}^2$. For a fixed dictionary $\bar{\boldsymbol{\mathcal{G}}}$, the population and empirical risks are

$$\mathrm{R}(\boldsymbol{\alpha} \mid \bar{\boldsymbol{\mathcal{G}}}) := \mathbb{E}_\pi[\ell_t(G_{\boldsymbol{\alpha}})], \qquad \widehat{\mathrm{R}}(\boldsymbol{\alpha} \mid \bar{\boldsymbol{\mathcal{G}}}) := \frac{1}{n}\sum_{t=1}^{n} \ell_t(G_{\boldsymbol{\alpha}}), \tag{106}$$

with minimizers $\boldsymbol{\alpha}^*(\bar{\boldsymbol{\mathcal{G}}}) \in \arg\min_{\boldsymbol{\alpha}} \mathrm{R}(\boldsymbol{\alpha} \mid \bar{\boldsymbol{\mathcal{G}}})$ and $\hat{\boldsymbol{\alpha}}(\bar{\boldsymbol{\mathcal{G}}}) \in \arg\min_{\boldsymbol{\alpha}} \widehat{\mathrm{R}}(\boldsymbol{\alpha} \mid \bar{\boldsymbol{\mathcal{G}}})$. We identify the following quantities:

- $\kappa := \sup_{x\in\mathcal{X}} \|\phi(x)\|_{\mathcal{H}}$, the uniform bound on the feature map;
- $M := \sup_{j\in[d]} \|\mathrm{Op}[\bar{\boldsymbol{\mathcal{G}}}_j]\|_{\mathrm{op}}$, the operator norm of the dictionary atoms; $\sup_{\boldsymbol{\alpha}} \|G_{\boldsymbol{\alpha}}\|_{\mathrm{op}} \leq M$ since SGOT barycenters do not increase operator norm;
- $m \leq d-1$, the intrinsic dimension of $\mathcal{G}(\bar{\boldsymbol{\mathcal{G}}})$ under $\|\cdot\|_{\mathrm{op}}$, with diameter $\leq 2M$;
- $\sigma^2_{\bar{\boldsymbol{\mathcal{G}}}} := \sup_{G\in\mathcal{G}(\bar{\boldsymbol{\mathcal{G}}})} \mathrm{Var}_\pi(\ell_t(G))$, the uniform variance of the sample loss;
- $\bar{\sigma}^2_{\mathcal{G}} := \sup_{G\in\mathcal{G}(\bar{\boldsymbol{\mathcal{G}}})} \mathbb{E}_\pi[\ell_t(G)^2]$, the uniform second moment of the sample loss.

F.2 Technical results

Regularity and entropy of the decoder

Lemma 3 (Proposition 5 of [17])**.** *For all $G, G' \in \mathcal{S}_r(\mathcal{H})$, $\|G - G'\|_{\mathrm{op}} \leq C_{\mathcal{S}}\, d_{\mathcal{S}}(G, G')$, where $C_{\mathcal{S}} = 2\sqrt{2}\, C_{\xi\psi}$.*

Lemma 4 (Lipschitz continuity of the OT barycenter decoder)**.** *Let $\bar{\boldsymbol{\mathcal{G}}} = (\bar{\boldsymbol{\mathcal{G}}}_j)_{j\in[d]} \in \mathcal{N}^d$ be a fixed dictionary with atoms of bounded operator norm $M < \infty$. Then $\boldsymbol{\alpha} \mapsto G_{\boldsymbol{\alpha}} = \mathrm{Op}[B^{\mathrm{OT}}(\boldsymbol{\alpha}; \bar{\boldsymbol{\mathcal{G}}})]$ is L_{dec}-Lipschitz from $(\Delta^{d-1}, \|\cdot\|_1)$ to $(\mathcal{S}_r(\mathcal{H}), \|\cdot\|_{\mathrm{op}})$, with $L_{\mathrm{dec}} := C_{\mathcal{S}} \max_{j,k\in[d]} d_{\mathcal{S}}(\mathrm{Op}[\bar{\boldsymbol{\mathcal{G}}}_j], \mathrm{Op}[\bar{\boldsymbol{\mathcal{G}}}_k])$.*

Proof. By stability of SGOT barycenters with respect to weights [17],

$$d_{\mathcal{S}}(G_{\boldsymbol{\alpha}}, G_{\boldsymbol{\alpha}'}) \leq \max_{j,k\in[d]} d_{\mathcal{S}}\big(\mathrm{Op}[\bar{\boldsymbol{\mathcal{G}}}_j],\, \mathrm{Op}[\bar{\boldsymbol{\mathcal{G}}}_k]\big) \cdot \|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_1. \tag{107}$$

Then $\boldsymbol{\alpha} \mapsto G_{\boldsymbol{\alpha}} = \mathrm{Op}[\mathrm{B}^{\mathrm{OT}}(\boldsymbol{\alpha}; \bar{\boldsymbol{\mathcal{G}}})]$ is Lipschitz from $(\Delta^{d-1}, \|\cdot\|_1)$ to $(\mathcal{S}_r(\mathcal{H}), d_{\mathcal{S}})$ with Lipschitz constant bounded from above by $\max_{j,k\in[d]} d_{\mathcal{S}}(\mathrm{Op}[\bar{\boldsymbol{\mathcal{G}}}_j], \mathrm{Op}[\bar{\boldsymbol{\mathcal{G}}}_k])$.

Applying Lemma 3, $\|G_{\boldsymbol{\alpha}} - G_{\boldsymbol{\alpha}'}\|_{\mathrm{op}} \leq C_{\mathcal{S}}\, d_{\mathcal{S}}(G_{\boldsymbol{\alpha}}, G_{\boldsymbol{\alpha}'}) \leq C_{\mathcal{S}} L_{\mathrm{dec}} \|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_1$. □

Lemma 5 (Metric entropy of the OT barycenter decoder image)**.** *Let $\mathcal{G}(\bar{\boldsymbol{\mathcal{G}}}) = \{G_{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \Delta^{d-1}\}$ for a fixed dictionary $\bar{\boldsymbol{\mathcal{G}}}$, and let $m \leq d-1$ be the intrinsic dimension of $\mathcal{G}(\bar{\boldsymbol{\mathcal{G}}})$ under $\|\cdot\|_{\mathrm{op}}$. Then for every $\varepsilon > 0$,*

$$\log \mathcal{N}\big(\mathcal{G}(\bar{\boldsymbol{\mathcal{G}}}),\, \|\cdot\|_{\mathrm{op}},\, \varepsilon\big) \;\lesssim\; m \log\left(\frac{L_{\mathrm{dec}}}{\varepsilon}\right). \tag{108}$$

Proof. Since $\mathcal{G}(\bar{\boldsymbol{\mathcal{G}}})$ has intrinsic dimension m under $\|\cdot\|_{\mathrm{op}}$, a standard volumetric argument gives $\log \mathcal{N}(\mathcal{G}(\bar{\boldsymbol{\mathcal{G}}}), \|\cdot\|_{\mathrm{op}}, \varepsilon) \lesssim m \log(D_{\mathcal{G}}/\varepsilon)$, where $D_{\mathcal{G}} \leq L_{\mathrm{dec}}$ is the diameter of $\mathcal{G}(\bar{\boldsymbol{\mathcal{G}}})$ under $\|\cdot\|_{\mathrm{op}}$, since $\mathcal{G}(\bar{\boldsymbol{\mathcal{G}}})$ is the image of Δ^{d-1} under an L_{dec}-Lipschitz map from $(\Delta^{d-1}, \|\cdot\|_1)$ of diameter 1. □

Lemma 6 (Lipschitz continuity of the projection-based decoder)**.** *Let $\bar{\mathcal{G}} = (\bar{\mathcal{G}}_j)_{j\in[d]} \in \mathcal{N}^d$ be a fixed dictionary with atoms $\bar{\mathcal{G}}_j = (\bar{\mathbf{\Lambda}}_j, \bar{\mathbf{L}}_j, \bar{\mathbf{R}}_j)$. Assume that the dictionary atoms have bounded operator norm $M < \infty$ and that the right eigenvector aggregation is uniformly non-degenerate:*

$$\sigma_{\min}\left(\sum_{j\in[d]} \alpha_j \bar{\mathbf{R}}_j\right) \geq c > 0, \qquad \forall\, \boldsymbol{\alpha} \in \Delta^{d-1}. \tag{109}$$

Then $\boldsymbol{\alpha} \mapsto G_{\boldsymbol{\alpha}} = Op[B^p(\boldsymbol{\alpha}; \bar{\mathcal{G}})]$ is Lipschitz from $(\Delta^{d-1}, \|\cdot\|_1)$ to $(\mathcal{S}_r(\mathcal{H}), \|\cdot\|_{\mathrm{op}})$ with constant

$$L_{\mathrm{dec}} \;\lesssim\; M^2(1+M^2)^2/c^4, \tag{110}$$

and in particular is $L_{\mathrm{dec}} C_{\mathcal{S}}$-Lipschitz from $(\Delta^{d-1}, \|\cdot\|_1)$ to $(\mathcal{S}_r(\mathcal{H}), d_{\mathcal{S}})$.

Proof. Fix $\boldsymbol{\alpha}, \boldsymbol{\alpha}' \in \Delta^{d-1}$ and write $\mathbf{R} := \sum_j \alpha_j \bar{\mathbf{R}}_j$, $\mathbf{R}' := \sum_j \alpha'_j \bar{\mathbf{R}}_j$, $\mathbf{L} := \sum_j \alpha_j \bar{\mathbf{L}}_j$, $\mathbf{L}' := \sum_j \alpha'_j \bar{\mathbf{L}}_j$, $\mathbf{\Lambda} := \sum_j \alpha_j \bar{\mathbf{\Lambda}}_j$, $\mathbf{\Lambda}' := \sum_j \alpha'_j \bar{\mathbf{\Lambda}}_j$. By linearity and $\|\cdot\|_1$-contractivity of convex combinations,

$$\|\mathbf{R} - \mathbf{R}'\|_{\mathrm{op}},\ \|\mathbf{L} - \mathbf{L}'\|_{\mathrm{op}},\ |\mathbf{\Lambda} - \mathbf{\Lambda}'| \;\leq\; M\|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_1. \tag{111}$$

Projection step. The projection $\mathrm{P}_{\mathcal{N}}$ maps $(\mathbf{\Lambda}, \mathbf{L}, \mathbf{R})$ to $(\mathbf{\Lambda}, \tilde{\mathbf{L}}, \mathbf{R})$ where

$$\tilde{\mathbf{L}} = \mathbf{L} + \mathbf{R}(\mathbf{R}^*\mathbf{R})^{-1}(\mathbf{I}_r - \mathbf{L}^*\mathbf{R})^*. \tag{112}$$

To see this, note that any $\tilde{\mathbf{L}}$ satisfying $\tilde{\mathbf{L}}^*\mathbf{R} = \mathbf{I}_r$ can be written as $\tilde{\mathbf{L}} = \mathbf{L} + \mathbf{R}(\mathbf{R}^*\mathbf{R})^{-1}\mathbf{V}^*$ for some matrix $\mathbf{V}$, and minimizing $\|\tilde{\mathbf{L}} - \mathbf{L}\|_{\mathrm{op}}$ over $\mathbf{V}$ gives $\mathbf{V} = \mathbf{I}_r - \mathbf{L}^*\mathbf{R}$, yielding (112). We bound $\|\tilde{\mathbf{L}} - \tilde{\mathbf{L}}'\|_{\mathrm{op}}$ by writing

$$\tilde{\mathbf{L}} - \tilde{\mathbf{L}}' = (\mathbf{L} - \mathbf{L}') + \mathbf{R}(\mathbf{R}^*\mathbf{R})^{-1}(\mathbf{I}_r - \mathbf{L}^*\mathbf{R})^* - \mathbf{R}'(\mathbf{R}'^*\mathbf{R}')^{-1}(\mathbf{I}_r - \mathbf{L}'^*\mathbf{R}')^*. \tag{113}$$

The first term satisfies $\|\mathbf{L} - \mathbf{L}'\|_{\mathrm{op}} \leq M\|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_1$ by (111). For the second term, decompose as

$$\begin{aligned}
&\mathbf{R}(\mathbf{R}^*\mathbf{R})^{-1}(\mathbf{I}_r - \mathbf{L}^*\mathbf{R})^* - \mathbf{R}'(\mathbf{R}'^*\mathbf{R}')^{-1}(\mathbf{I}_r - \mathbf{L}'^*\mathbf{R}')^* \\
&= \big[\mathbf{R}(\mathbf{R}^*\mathbf{R})^{-1} - \mathbf{R}'(\mathbf{R}'^*\mathbf{R}')^{-1}\big](\mathbf{I}_r - \mathbf{L}^*\mathbf{R})^* + \mathbf{R}'(\mathbf{R}'^*\mathbf{R}')^{-1}\big[(\mathbf{L}'^*\mathbf{R}')^* - (\mathbf{L}^*\mathbf{R})^*\big].
\end{aligned} \tag{114}$$

First bracket. Using $\mathbf{R}(\mathbf{R}^*\mathbf{R})^{-1} - \mathbf{R}'(\mathbf{R}'^*\mathbf{R}')^{-1} = (\mathbf{R} - \mathbf{R}')(\mathbf{R}^*\mathbf{R})^{-1} + \mathbf{R}'[(\mathbf{R}^*\mathbf{R})^{-1} - (\mathbf{R}'^*\mathbf{R}')^{-1}]$, the resolvent identity, $\|(\mathbf{R}^*\mathbf{R})^{-1}\|_{\mathrm{op}} \leq c^{-2}$, and $\|\mathbf{R}'^*\mathbf{R}' - \mathbf{R}^*\mathbf{R}\|_{\mathrm{op}} \leq 2M\|\mathbf{R} - \mathbf{R}'\|_{\mathrm{op}}$,

$$\|\mathbf{R}(\mathbf{R}^*\mathbf{R})^{-1} - \mathbf{R}'(\mathbf{R}'^*\mathbf{R}')^{-1}\|_{\mathrm{op}} \;\leq\; \frac{M(1+2M^2/c^2)}{c^2}\|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_1. \tag{115}$$

The factor $\|(\mathbf{I}_r - \mathbf{L}^*\mathbf{R})^*\|_{\mathrm{op}} \leq 1 + \|\mathbf{L}\|_{\mathrm{op}}\|\mathbf{R}\|_{\mathrm{op}} \leq 1 + M^2$.

Second bracket. $\|\mathbf{L}'^*\mathbf{R}' - \mathbf{L}^*\mathbf{R}\|_{\mathrm{op}} \leq \|\mathbf{L}' - \mathbf{L}\|_{\mathrm{op}}\|\mathbf{R}'\|_{\mathrm{op}} + \|\mathbf{L}\|_{\mathrm{op}}\|\mathbf{R}' - \mathbf{R}\|_{\mathrm{op}} \leq 2M^2\|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_1$, and $\|\mathbf{R}'(\mathbf{R}'^*\mathbf{R}')^{-1}\|_{\mathrm{op}} \leq M/c^2$.

Combining both brackets,

$$\|\tilde{\mathbf{L}} - \tilde{\mathbf{L}}'\|_{\mathrm{op}} \;\lesssim\; \frac{M(1+M^2)^2}{c^4}\|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_1. \tag{116}$$

Operator norm step. Writing

$$G_{\boldsymbol{\alpha}} - G_{\boldsymbol{\alpha}'} = (\mathbf{R} - \mathbf{R}')\mathrm{Diag}(\mathbf{\Lambda})\tilde{\mathbf{L}}^* + \mathbf{R}'\mathrm{Diag}(\mathbf{\Lambda} - \mathbf{\Lambda}')\tilde{\mathbf{L}}^* + \mathbf{R}'\mathrm{Diag}(\mathbf{\Lambda}')(\tilde{\mathbf{L}} - \tilde{\mathbf{L}}')^*, \tag{117}$$

and using $\|\mathbf{R}\|_{\mathrm{op}}, \|\mathbf{R}'\|_{\mathrm{op}} \leq M$, $|\mathbf{\Lambda}|, |\mathbf{\Lambda}'| \leq M$, $\|\tilde{\mathbf{L}}\|_{\mathrm{op}} \leq M + M(1+M^2)/c^2$, together with (111) and (116),

$$\|G_{\boldsymbol{\alpha}} - G_{\boldsymbol{\alpha}'}\|_{\mathrm{op}} \;\lesssim\; \frac{M^2(1+M^2)^2}{c^4}\|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_1. \tag{118}$$

This establishes Lipschitz continuity with constant $L_{\mathrm{dec}} \lesssim M^2(1+M^2)^2/c^4$. □

Lemma 7 (Metric entropy of the projection-based decoder image, B^{p})**.** *Let Assumption* **(WC)** *be satisfied. Then for every $\varepsilon > 0$,*

$$\log \mathcal{N}\big(\mathcal{G}(\bar{\boldsymbol{\mathcal{G}}}),\, \|\cdot\|_{\mathrm{op}},\, \varepsilon\big) \;\lesssim\; m \log\!\left(\frac{M}{c^2\,\varepsilon}\right), \tag{119}$$

where $m \le d-1$ is the intrinsic dimension of $\mathcal{G}(\bar{\boldsymbol{\mathcal{G}}})$ under $\|\cdot\|_{\mathrm{op}}$.

Proof. Under Assumption **(WC)**, the map $\boldsymbol{\alpha} \mapsto G_{\boldsymbol{\alpha}} = \mathrm{Op}[\mathrm{B}^{\mathrm{p}}(\boldsymbol{\alpha}; \bar{\boldsymbol{\mathcal{G}}})]$ is L_{dec}-Lipschitz from $(\Delta^{d-1}, \|\cdot\|_1)$ to $(\mathcal{S}_r(\mathcal{H}), \|\cdot\|_{\mathrm{op}})$ with $L_{\mathrm{dec}} \lesssim M(1+M/c^2)$ by Lemma 6. Note also that the diameter satisfies $D_{\mathcal{G}} \le 2M$. The result now follows from an identical argument to that of Lemma 5. □

Oracle dictionary and dictionary bias. We consider d training dynamical systems, each observed through a trajectory of length n_{train}, from which the transfer operators $\{\hat{\boldsymbol{\mathcal{G}}}_j\}_{j\in[d]}$ are estimated. The empirically learned dictionary $\hat{\boldsymbol{\mathcal{G}}}$ minimizes objective (4) over these estimated operators. The oracle dictionary $\bar{\boldsymbol{\mathcal{G}}}^*$ is the minimizer of the same objective evaluated at the true operators $\{\boldsymbol{\mathcal{G}}_j\}_{j\in[d]}$, i.e. what dictionary learning would produce from infinitely long training trajectories:

$$\bar{\boldsymbol{\mathcal{G}}}^* \in \operatorname*{arg\,min}_{\bar{\boldsymbol{\mathcal{G}}}\in\mathcal{N}^d} \; \frac{1}{d}\sum_{j=1}^{d} \min_{\boldsymbol{\alpha}\in\Delta^{d-1}} d_{\mathcal{S}}\big(\mathrm{B}(\boldsymbol{\alpha}; \bar{\boldsymbol{\mathcal{G}}}),\, \boldsymbol{\mathcal{G}}_j\big)\,. \tag{120}$$

The dictionary bias

$$\mathrm{bias}(\hat{\boldsymbol{\mathcal{G}}}) := \inf_{\boldsymbol{\alpha}\in\Delta^{d-1}} \mathrm{R}(\boldsymbol{\alpha} \mid \hat{\boldsymbol{\mathcal{G}}}) - \inf_{\boldsymbol{\alpha}\in\Delta^{d-1}} \mathrm{R}(\boldsymbol{\alpha} \mid \bar{\boldsymbol{\mathcal{G}}}^*) \;\ge\; 0 \tag{121}$$

measures the effect of operator estimation error on the learned dictionary. It vanishes when $n_{\mathrm{train}} \to \infty$. For two dictionaries $\bar{\boldsymbol{\mathcal{G}}}, \bar{\boldsymbol{\mathcal{G}}}' \in \mathcal{N}^d$, we extend the operator norm componentwise by

$$\|\hat{\boldsymbol{\mathcal{G}}} - \bar{\boldsymbol{\mathcal{G}}}^*\|_{\mathrm{op}} := \max_{j\in[d]} \|\mathrm{Op}[\hat{\boldsymbol{\mathcal{G}}}_j] - \mathrm{Op}[\bar{\boldsymbol{\mathcal{G}}}^*_j]\|_{\mathrm{op}}. \tag{122}$$

Proposition 7 (Control of the dictionary bias)**.** *Under the L_{dec}-Lipschitz assumption on the decoder in $\bar{\boldsymbol{\mathcal{G}}}$ under $\|\cdot\|_{\mathrm{op}}$ (Lemmas 4 and 6),*

$$\mathrm{bias}(\hat{\boldsymbol{\mathcal{G}}}) \;\le\; 2(1+M)\kappa^2 L_{\mathrm{dec}}\, \|\hat{\boldsymbol{\mathcal{G}}} - \bar{\boldsymbol{\mathcal{G}}}^*\|_{\mathrm{op}}. \tag{123}$$

Proof. For any $\boldsymbol{\alpha} \in \Delta^{d-1}$ and two dictionaries $\bar{\boldsymbol{\mathcal{G}}}, \bar{\boldsymbol{\mathcal{G}}}' \in \mathcal{N}^d$, let $G_{\boldsymbol{\alpha}} = \mathrm{Op}[\mathrm{B}(\boldsymbol{\alpha}; \bar{\boldsymbol{\mathcal{G}}})]$ and $G'_{\boldsymbol{\alpha}} = \mathrm{Op}[\mathrm{B}(\boldsymbol{\alpha}; \bar{\boldsymbol{\mathcal{G}}}')]$. Writing $e_t(G) := \mathrm{z}_{t+1} - G^*\mathrm{z}_t$,

$$|\ell_t(G) - \ell_t(\bar{G})| = \big|\langle e_t(G) + e_t(\bar{G}), (G-\bar{G})^*\mathrm{z}_t\rangle_{\mathcal{H}}\big| \le \big(\|e_t(G)\|_{\mathcal{H}} + \|e_t(\bar{G})\|_{\mathcal{H}}\big)\|G-\bar{G}\|_{\mathrm{op}}\|\mathrm{z}_t\|_{\mathcal{H}}. \tag{124}$$

Since $\|\mathrm{z}_t\|_{\mathcal{H}} \le \kappa$ and $\|G\|_{\mathrm{op}}, \|\bar{G}\|_{\mathrm{op}} \le M$, we have $\|e_t(G)\|_{\mathcal{H}} \le (1+M)\kappa$. The L_{dec}-Lipschitz assumption on the decoder in $\|\cdot\|_{\mathrm{op}}$,

$$|\mathrm{R}(\boldsymbol{\alpha} \mid \bar{\boldsymbol{\mathcal{G}}}) - \mathrm{R}(\boldsymbol{\alpha} \mid \bar{\boldsymbol{\mathcal{G}}}')| \le \mathbb{E}_\pi|\ell_t(G_{\boldsymbol{\alpha}}) - \ell_t(G'_{\boldsymbol{\alpha}})| \le 2(1+M)\kappa^2 L_{\mathrm{dec}}\, \|\bar{\boldsymbol{\mathcal{G}}} - \bar{\boldsymbol{\mathcal{G}}}'\|_{\mathrm{op}}. \tag{125}$$

The result follows from the value function inequality $|\inf_{\boldsymbol{\alpha}} f(\boldsymbol{\alpha}) - \inf_{\boldsymbol{\alpha}} g(\boldsymbol{\alpha})| \le \sup_{\boldsymbol{\alpha}} |f(\boldsymbol{\alpha}) - g(\boldsymbol{\alpha})|$. □

Remark 5 (Rate of the dictionary bias)**.** *By Lipschitz continuity of the dictionary learning objective* (4) *in the training operators, $\|\hat{\boldsymbol{\mathcal{G}}} - \bar{\boldsymbol{\mathcal{G}}}^*\|_{\mathrm{op}} \lesssim \max_{j\in[d]} \|Op[\hat{\boldsymbol{\mathcal{G}}}_j] - Op[\boldsymbol{\mathcal{G}}_j]\|_{\mathrm{op}}$. Each operator estimation error $\|Op[\hat{\boldsymbol{\mathcal{G}}}_j] - Op[\boldsymbol{\mathcal{G}}_j]\|_{\mathrm{op}}$ is controlled by the spectral estimation rates of Appendix A.1, see* (53)*, with trajectory length n_{train}. A union bound over $j \in [d]$ gives with probability at least $1-\delta$,*

$$\mathrm{bias}(\hat{\boldsymbol{\mathcal{G}}}) \;\lesssim\; (1+M)\kappa^2 L_{\mathrm{dec}}\, (n_{\mathrm{train}})^{-\frac{\alpha}{2(\alpha+\beta)}} \log(d/\delta), \tag{126}$$

where $\alpha \in (0,2]$ and $\beta \in (0,1]$ are the regularity and spectral decay parameters of Appendix A.1. The bias decays with the training trajectory length n_{train}, confirming that longer training trajectories yield a learned dictionary closer to the oracle.

F.3 Main results

Coordinate fitting guarantee.

Theorem 2 (Coordinate fitting, i.i.d.)**.** *Let $\bar{\boldsymbol{\mathcal{G}}} \in \mathcal{N}^d$ be a fixed dictionary. Suppose $(X_t)_{t\in[n]}$ has marginal π and is i.i.d. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\mathrm{R}(\hat{\boldsymbol{\alpha}} \mid \bar{\boldsymbol{\mathcal{G}}}) - \inf_{\boldsymbol{\alpha}\in\Delta^{d-1}} \mathrm{R}(\boldsymbol{\alpha} \mid \bar{\boldsymbol{\mathcal{G}}}) \;\le\; \mathrm{bias}(\hat{\boldsymbol{\mathcal{G}}}) + c' \sqrt{\frac{\sigma^2_{\bar{\boldsymbol{\mathcal{G}}}} \cdot m \log(M L_{\mathrm{dec}}(1+M)\kappa^2\sqrt{n}) + \log(1/\delta)}{n}}, \tag{127}$$

where $c' > 0$ is a numerical constant.

Proof of Theorem 2. We proceed in five steps.

Step 1: Reduction to uniform deviation. Decompose the excess risk as

$$\mathrm{R}(\hat{\boldsymbol{\alpha}} \mid \hat{\boldsymbol{\mathcal{G}}}) - \inf_{\boldsymbol{\alpha}\in\Delta^{d-1}} \mathrm{R}(\boldsymbol{\alpha} \mid \bar{\boldsymbol{\mathcal{G}}}^*) = \underbrace{\mathrm{R}(\hat{\boldsymbol{\alpha}} \mid \hat{\boldsymbol{\mathcal{G}}}) - \inf_{\boldsymbol{\alpha}} \mathrm{R}(\boldsymbol{\alpha} \mid \hat{\boldsymbol{\mathcal{G}}})}_{\text{coordinate fitting excess risk}} + \underbrace{\inf_{\boldsymbol{\alpha}} \mathrm{R}(\boldsymbol{\alpha} \mid \hat{\boldsymbol{\mathcal{G}}}) - \inf_{\boldsymbol{\alpha}} \mathrm{R}(\boldsymbol{\alpha} \mid \bar{\boldsymbol{\mathcal{G}}}^*)}_{\mathrm{bias}(\hat{\boldsymbol{\mathcal{G}}})}. \tag{128}$$

We now bound the coordinate fitting excess risk. By optimality of $\hat{\boldsymbol{\alpha}}$ for $\widehat{\mathrm{R}}$ and of $\boldsymbol{\alpha}^*(\hat{\boldsymbol{\mathcal{G}}})$ for $\mathrm{R}(\cdot \mid \hat{\boldsymbol{\mathcal{G}}})$,

$$\mathrm{R}(\hat{\boldsymbol{\alpha}} \mid \hat{\boldsymbol{\mathcal{G}}}) - \inf_{\boldsymbol{\alpha}} \mathrm{R}(\boldsymbol{\alpha} \mid \hat{\boldsymbol{\mathcal{G}}}) \le 2 \sup_{G\in\mathcal{G}(\hat{\boldsymbol{\mathcal{G}}})} \big|\bar{\ell}(G) - \mathbb{E}_\pi[\ell_t(G)]\big|, \tag{129}$$

where $\bar{\ell}(G) := \frac{1}{n}\sum_{t=1}^n \ell_t(G)$.

Step 2: Covering net. For $\varepsilon > 0$ to be chosen, let $\mathcal{G}_\varepsilon$ be an ε-net of $\mathcal{G}(\bar{\boldsymbol{\mathcal{G}}})$ under $\|\cdot\|_{\mathrm{op}}$. By Lemmas 4 and 6 and the volumetric argument of Lemma 5, $\log|\mathcal{G}_\varepsilon| \lesssim m\log(ML_{\mathrm{dec}}/\varepsilon)$. For any $G \in \mathcal{G}(\bar{\boldsymbol{\mathcal{G}}})$, pick $\bar{G} \in \mathcal{G}_\varepsilon$ with $\|G - \bar{G}\|_{\mathrm{op}} \le \varepsilon$.

Step 3: Lipschitz continuity of the loss. Writing $e_t(G) := \mathrm{z}_{t+1} - G^*\mathrm{z}_t$,

$$|\ell_t(G) - \ell_t(\bar{G})| = \big|\langle e_t(G) + e_t(\bar{G}), (G-\bar{G})^*\mathrm{z}_t\rangle_{\mathcal{H}}\big| \le \big(\|e_t(G)\|_{\mathcal{H}} + \|e_t(\bar{G})\|_{\mathcal{H}}\big)\|G-\bar{G}\|_{\mathrm{op}}\|\mathrm{z}_t\|_{\mathcal{H}}. \tag{130}$$

Since $\|\mathrm{z}_t\|_{\mathcal{H}} \le \kappa$ and $\|G\|_{\mathrm{op}}, \|\bar{G}\|_{\mathrm{op}} \le M$, we have $\|e_t(G)\|_{\mathcal{H}} \le (1+M)\kappa$, and thus

$$|\ell_t(G) - \ell_t(\bar{G})| \le 2(1+M)\kappa^2 \,\|G-\bar{G}\|_{\mathrm{op}} \le 2(1+M)\kappa^2\,\varepsilon. \tag{131}$$

Step 4: Bernstein's inequality. For fixed $\bar{G} \in \mathcal{G}_\varepsilon$, $0 \le \ell_t(\bar{G}) \le (1+M)^2\kappa^2$ and $\mathrm{Var}_\pi(\ell_t(\bar{G})) \le \sigma^2_{\bar{\boldsymbol{\mathcal{G}}}}$. By Bernstein's inequality applied to the i.i.d. sequence $(\ell_t(\bar{G}))_{t=1}^n$, for any $s > 0$,

$$\mathbb{P}\Big(\big|\bar{\ell}(\bar{G}) - \mathbb{E}_\pi[\ell_t(\bar{G})]\big| \ge \frac{s}{2}\Big) \le 2\exp\left(-\frac{n\,s^2/4}{2\sigma^2_{\bar{\boldsymbol{\mathcal{G}}}} + \frac{2}{3}(1+M)^2\kappa^2\,s}\right). \tag{132}$$

Step 5: Union bound and optimization. Set $\varepsilon := s/(4(1+M)\kappa^2)$ so that $2(1+M)\kappa^2\varepsilon = s/2$. By the triangle inequality, $|\bar{\ell}(G) - \mathbb{E}_\pi[\ell_t(G)]| \le |\bar{\ell}(\bar{G}) - \mathbb{E}_\pi[\ell_t(\bar{G})]| + s/2$. The net size at this ε satisfies $\log|\mathcal{G}_\varepsilon| \lesssim m\log(ML_{\mathrm{dec}}(1+M)\kappa^2/s)$. Taking a union bound over $\mathcal{G}_\varepsilon$ and solving for s such that the right-hand side equals δ yields

$$\sup_{G\in\mathcal{G}(\bar{\boldsymbol{\mathcal{G}}})} \big|\bar{\ell}(G) - \mathbb{E}_\pi[\ell_t(G)]\big| \;\lesssim\; \sqrt{\frac{\sigma^2_{\bar{\boldsymbol{\mathcal{G}}}} \cdot m\log(ML_{\mathrm{dec}}(1+M)\kappa^2\sqrt{n}) + \log(1/\delta)}{n}}. \tag{133}$$

The result follows from Step 1. □

Extension to β-mixing trajectories. For a stationary process $(X_t)_{t\geq 1}$ with marginal π, the β-mixing coefficients are

$$\beta(\tau) := \sup_{t\geq 1} \sup_{B\in\sigma(X_s,\, s\leq t)\otimes\sigma(X_s,\, s\geq t+\tau)} \big|\mathbb{P}(B) - (\mathbb{P}\otimes\mathbb{P})(B)\big|, \quad \tau\in\mathbb{N}. \tag{134}$$

Fix block length $\tau\in\mathbb{N}$, let $n_{\mathrm{eff}} := \lfloor n/(2\tau)\rfloor$, and define the effective variance

$$\bar{\sigma}^2_{\mathrm{eff}} := \sigma^2_{\bar{\boldsymbol{\mathcal{G}}}} + 4\,\bar{\sigma}^2_{\mathcal{G}} \sum_{s=1}^{\tau-1} \beta(s), \tag{135}$$

which accounts for within-block temporal dependence via the Davydov–Rio covariance inequality [52].

Theorem 1 (Oracle inequality)**.** *Let* **(WC)** *be satisfied. Suppose the test trajectory $(X_t)_{t\in[n+1]}$ is stationary, β-mixing, and independent of the training data. Then for any $\delta\in(0,1)$, with probability at least $1-\delta-2(n_{\mathrm{eff}}-1)\beta(\tau)$,*

$$\mathrm{R}(\hat{\boldsymbol{\alpha}} \mid \hat{\boldsymbol{\mathcal{G}}}) - \inf_{\boldsymbol{\alpha}} \mathrm{R}(\boldsymbol{\alpha} \mid \bar{\boldsymbol{\mathcal{G}}}) \;\leq \mathrm{bias}(\hat{\boldsymbol{\mathcal{G}}}) + c' \sqrt{\frac{\bar{\sigma}^2_{\mathrm{eff}} \cdot m \log\big(M L_{\mathrm{dec}}(1+M)\kappa^2\sqrt{n_{\mathrm{eff}}}\big) + \log(1/\delta)}{n_{\mathrm{eff}}}} \tag{8}$$

where $c'>0$ is a numerical constant and $\mathrm{bias}(\hat{\boldsymbol{\mathcal{G}}}) := \inf_{\boldsymbol{\alpha}} \mathrm{R}(\boldsymbol{\alpha} \mid \hat{\boldsymbol{\mathcal{G}}}) - \inf_{\boldsymbol{\alpha}} \mathrm{R}(\boldsymbol{\alpha} \mid \bar{\boldsymbol{\mathcal{G}}}^*) \geq 0$.

Proof. **Steps 1–3: Reduction, covering, and Lipschitz control.** By optimality of $\hat{\boldsymbol{\alpha}}$,

$$\mathrm{R}(\hat{\boldsymbol{\alpha}} \mid \hat{\boldsymbol{\mathcal{G}}}) - \inf_{\boldsymbol{\alpha}} \mathrm{R}(\boldsymbol{\alpha} \mid \hat{\boldsymbol{\mathcal{G}}}) \leq 2 \sup_{G\in\mathcal{G}(\hat{\boldsymbol{\mathcal{G}}})} \big|\bar{\ell}(G) - \mathbb{E}_\pi[\ell_t(G)]\big|. \tag{136}$$

For $\varepsilon>0$ to be chosen, let $\mathcal{G}_\varepsilon$ be an ε-net of $\mathcal{G}(\hat{\boldsymbol{\mathcal{G}}})$ under $\|\cdot\|_{\mathrm{op}}$, with $\log|\mathcal{G}_\varepsilon| \lesssim m\log(ML_{\mathrm{dec}}/\varepsilon)$. For any $G\in\mathcal{G}(\hat{\boldsymbol{\mathcal{G}}})$, pick $\bar{G}\in\mathcal{G}_\varepsilon$ with $\|G-\bar{G}\|_{\mathrm{op}}\leq\varepsilon$. By Step 3 of Theorem 2, $|\ell_t(G)-\ell_t(\bar{G})| \leq 2(1+M)\kappa^2\,\varepsilon$.

Step 4: Control of $\mathrm{Var}(Y_j)$. Following the blocking construction of [34, Lemma 1], split $(\ell_t(\bar{G}))_{t=1}^n$ into odd block sums $Y_j := \sum_{i=2(j-1)\tau+1}^{(2j-1)\tau} \ell_{t_i}(\bar{G})$, $j=1,\dots,n_{\mathrm{eff}}$. Expanding the variance of Y_j,

$$\mathrm{Var}(Y_j) = \sum_{i=1}^{\tau} \mathrm{Var}(\ell_{t_i}(\bar{G})) + 2 \sum_{1\leq i<k\leq\tau} \mathrm{Cov}\big(\ell_{t_i}(\bar{G}),\, \ell_{t_k}(\bar{G})\big)\,. \tag{137}$$

The diagonal terms satisfy $\sum_{i=1}^{\tau}\mathrm{Var}(\ell_{t_i}(\bar{G})) \leq \tau\sigma^2_{\bar{\boldsymbol{\mathcal{G}}}}$ by stationarity. For the off-diagonal terms at lag $s=k-i$, the Davydov–Rio inequality [52] gives $|\mathrm{Cov}(\ell_{t_i}(\bar{G}),\ell_{t_k}(\bar{G}))| \leq 2\beta(s)\bar{\sigma}^2_{\mathcal{G}}$. Since lag $s\in\{1,\dots,\tau-1\}$ appears at most τ times,

$$\mathrm{Var}(Y_j) \leq \tau\sigma^2_{\bar{\boldsymbol{\mathcal{G}}}} + 4\tau\bar{\sigma}^2_{\mathcal{G}} \sum_{s=1}^{\tau-1}\beta(s) = \tau\bar{\sigma}^2_{\mathrm{eff}}. \tag{138}$$

Step 5: Bernstein for β-mixing.

The odd blocks are $\beta(\tau)$-approximately independent: replacing them by exactly independent copies incurs total variation error $2(n_{\mathrm{eff}}-1)\beta(\tau)$, following [34, Lemma 1]. Each Y_j satisfies $|Y_j - \mathbb{E}[Y_j]| \leq \tau(1+M)^2\kappa^2$ almost surely and $\mathrm{Var}(Y_j)\leq\tau\bar{\sigma}^2_{\mathrm{eff}}$ by (138). Applying Bernstein's inequality to the approximately independent odd blocks, for any $s>0$,

$$\mathbb{P}\left(\Big|\frac{1}{n_{\mathrm{eff}}}\sum_{j=1}^{n_{\mathrm{eff}}} Y_j - \mathbb{E}_\pi[\ell_t(\bar{G})]\Big| \geq \frac{s}{2}\right) \leq 2\exp\left(-\frac{n_{\mathrm{eff}}\,s^2/4}{2\bar{\sigma}^2_{\mathrm{eff}} + \frac{2}{3}(1+M)^2\kappa^2\,s}\right) + 2(n_{\mathrm{eff}}-1)\beta(\tau). \tag{139}$$

Step 6: Union bound and optimization.

Set $\varepsilon = s/(4(1+M)\kappa^2)$ so that $2(1+M)\kappa^2\varepsilon = s/2$. By the triangle inequality, $|\bar{\ell}(G)-\mathbb{E}_\pi[\ell_t(G)]| \leq |\bar{\ell}(\bar{G}) - \mathbb{E}_\pi[\ell_t(\bar{G})]| + s/2$. The net size at this ε satisfies $\log|\mathcal{G}_\varepsilon| \lesssim m\log(ML_{\mathrm{dec}}(1+M)\kappa^2/s)$. Taking a union bound over $\mathcal{G}_\varepsilon$ and solving for s such that the right-hand side equals δ yields the result, with the mixing tail $2(n_{\mathrm{eff}}-1)\beta(\tau)$ carried additively in the probability statement.

□

Remark 6 (Exponentially fast mixing)**.** *When $\beta(\tau) \le Ce^{-c\tau}$ for constants $C, c > 0$, the sum $\sum_{s=1}^{\tau-1} \beta(s) \le C/c$ is bounded uniformly in τ, so $\bar{\sigma}^2_{\text{eff}} \lesssim \sigma^2_{\bar{\mathcal{G}}} + \bar{\sigma}^2_{\mathcal{G}}$. Choosing $\tau = \lceil c^{-1} \log(n/\delta) \rceil$ ensures $2(n_{\text{eff}} - 1)\beta(\tau) \le \delta$, giving $n_{\text{eff}} = \Omega(n/\log(n/\delta))$ and recovering nearly the i.i.d. rate of Theorem 2.*

Remark 7 (Finite-dimensional embeddings)**.** *When ϕ takes values in $\mathbb{R}^p$, $\kappa = \sup_x \|\phi(x)\|_2 < \infty$ holds automatically with bounded activations such as* $\arctan$*, and $\|\cdot\|_{\mathcal{H}} = \|\cdot\|_2$ throughout. Both theorems apply without modification.*