Koopman Operator Regression: Statistical Learning Perspective to Data-driven Dynamical System

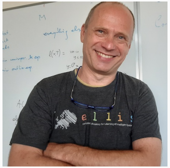
Vladimir Kostic

November 23, 2023

CSML, Italian Institute of Technology, Genoa, Italy
Dept. of Mathematics and Informatics, University of Novi Sad, Serbia

Papers:

- VK, P. Novelli, A. Maurer, C. Ciliberto, L. Rosasco, & M. Pontil. Learning dynamical systems via Koopman operator regression in RKHS. *NeurIPS 2022*
- VK, K. Lounici, P. Novelli & M. Pontil. Koopman Operator Learning: Sharp Spectral Rates and Spurious Eigenvalues *NeurIPS 2023*
- G. Meanti, A. Chatalic, VK, P. Novelli, M. Pontil & L. Rosasco. Estimating Koopman operators with sketching to provably learn large scale dynamical systems *NeurIPS 2023*
- VK, P. Novelli, R. Grazzi, K. Lounici & M. Pontil. Deep projection networks for learning time-homogeneous dynamical systems 2023



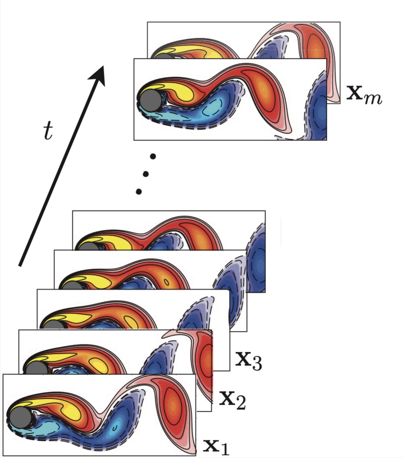
Plan

- Problem and Koopman operator approach
- Statistical learning formulation
- Numerical experiments
- Some Applications
- ERM and learning bounds
- Open problems

Problem & Our Approach

Problem

We wish to learn a dynamical system from data (trajectories) in a form that can:



- predict future states
- explain complex dynamics via recurring patterns
- interpret spacial and temporal relations of the states
- be used to control the dynamical process
- ...

An Easy Example: Noisy Linear Dynamics

State space $\mathcal{X} = \mathbb{R}^d$, $F \in \mathbb{R}^{d\times d}$ and $X_{t+1} = FX_t + \omega_t$, ω_t i.i.d. $\mathcal{N}(0, \sigma^2 I_d)$

An Easy Example: Noisy Linear Dynamics

State space $\mathcal{X} = \mathbb{R}^d$, $F \in \mathbb{R}^{d \times d}$ and $X_{t+1} = FX_t + \omega_t$, ω_t i.i.d. $\mathcal{N}(0, \sigma^2 I_d)$

- Eigenvalue decomposition (not an SVD!): $(\lambda_i, u_i, v_i) \in \mathbb{C} \times \mathbb{C}^d \times \mathbb{C}^d$,

$$Fv_i = \lambda_i v_i, \; u_i^* F = \lambda_i u_i^* \text{ and } u_i^* v_j = \delta_{ij} \quad \Longrightarrow \quad F = \sum_{i \in [d]} \lambda_i v_i u_i^*$$

An Easy Example: Noisy Linear Dynamics

State space $\mathcal{X} = \mathbb{R}^d$, $F \in \mathbb{R}^{d\times d}$ and $X_{t+1} = FX_t + \omega_t$, ω_t i.i.d. $\mathcal{N}(0, \sigma^2 I_d)$

- Eigenvalue decomposition (not an SVD!): $(\lambda_i, u_i, v_i) \in \mathbb{C} \times \mathbb{C}^d \times \mathbb{C}^d$,

$$Fv_i = \lambda_i v_i, \; u_i^* F = \lambda_i u_i^* \;\text{ and }\; u_i^* v_j = \delta_{ij} \quad \Longrightarrow \quad F = \sum_{i\in[d]} \lambda_i v_i u_i^*$$

- Expected dynamics: $\mathbb{E}[X_t \mid X_0 = x] = F^t x = \sum_{i\in[d]} \lambda_i^t \,(u_i^* x)\, v_i$

An Easy Example: Noisy Linear Dynamics

State space $\mathcal{X} = \mathbb{R}^d$, $F \in \mathbb{R}^{d\times d}$ and $X_{t+1} = FX_t + \omega_t$, ω_t i.i.d. $\mathcal{N}(0, \sigma^2 I_d)$

- Eigenvalue decomposition (not an SVD!): $(\lambda_i, u_i, v_i) \in \mathbb{C} \times \mathbb{C}^d \times \mathbb{C}^d$,

$$Fv_i = \lambda_i v_i,\ u_i^* F = \lambda_i u_i^* \ \text{ and }\ u_i^* v_j = \delta_{ij} \quad \Longrightarrow \quad F = \sum_{i\in[d]} \lambda_i v_i u_i^*$$

- Expected dynamics: $\mathbb{E}[X_t \mid X_0 = x] = F^t x = \sum_{i\in[d]} \lambda_i^t \, (u_i^* x) \, v_i$

A different perspective via measurements/observables :

An Easy Example: Noisy Linear Dynamics

State space $\mathcal{X} = \mathbb{R}^d$, $F \in \mathbb{R}^{d\times d}$ and $X_{t+1} = FX_t + \omega_t$, ω_t i.i.d. $\mathcal{N}(0, \sigma^2 I_d)$

- Eigenvalue decomposition (not an SVD!): $(\lambda_i, u_i, v_i) \in \mathbb{C} \times \mathbb{C}^d \times \mathbb{C}^d$,
$$F v_i = \lambda_i v_i,\ u_i^* F = \lambda_i u_i^* \ \text{ and } \ u_i^* v_j = \delta_{ij} \quad \Longrightarrow \quad F = \sum_{i\in[d]} \lambda_i v_i u_i^*$$
- Expected dynamics: $\mathbb{E}[X_t \,|\, X_0 = x] = F^t x = \sum_{i\in[d]} \lambda_i^t \,(u_i^* x)\, v_i$

A different perspective via measurements/observables :

- $\mathcal{F} := \{f_w := \langle \cdot, w\rangle : \mathbb{R}^d \to \mathbb{R} \,|\, w \in \mathbb{R}^d\} \implies \mathbb{E}[f_w(X_{t+1}) \,|\, X_t = x] = \langle x, F^* w\rangle = f_{F^* w}(x)$

An Easy Example: Noisy Linear Dynamics

State space $\mathcal{X} = \mathbb{R}^d$, $F \in \mathbb{R}^{d\times d}$ and $X_{t+1} = FX_t + \omega_t$, ω_t i.i.d. $\mathcal{N}(0, \sigma^2 I_d)$

- Eigenvalue decomposition (not an SVD!): $(\lambda_i, u_i, v_i) \in \mathbb{C} \times \mathbb{C}^d \times \mathbb{C}^d$,

$$F v_i = \lambda_i v_i,\; u_i^* F = \lambda_i u_i^* \;\text{ and }\; u_i^* v_j = \delta_{ij} \quad \Longrightarrow \quad F = \sum_{i\in[d]} \lambda_i v_i u_i^*$$

- Expected dynamics: $\mathbb{E}[X_t \,|\, X_0 = x] = F^t x = \sum_{i\in[d]} \lambda_i^t \,(u_i^* x)\, v_i$

A different perspective via measurements/observables :

- $\mathcal{F} := \{f_w := \langle \cdot, w\rangle : \mathbb{R}^d \to \mathbb{R} \,|\, w \in \mathbb{R}^d\} \;\Longrightarrow\; \mathbb{E}[f_w(X_{t+1}) \,|\, X_t = x] = \langle x, F^* w\rangle = f_{F^* w}(x)$

- Expected dynamics of observables :

$$f_w \in \mathcal{F} \;\Longrightarrow\; \mathbb{E}[f_w(X_t) \,|\, X_0 = x] = \langle x, (F^*)^t w\rangle = \sum_{i\in[d]} \lambda_i^t \,\langle f_w, f_{v_i}\rangle_{\mathcal{F}}\, f_{u_i}(x)$$

An Easy Example: Noisy Linear Dynamics

State space $\mathcal{X} = \mathbb{R}^d$, $F \in \mathbb{R}^{d\times d}$ and $X_{t+1} = F X_t + \omega_t$, ω_t i.i.d. $\mathcal{N}(0, \sigma^2 I_d)$

- Eigenvalue decomposition (not an SVD!): $(\lambda_i, u_i, v_i) \in \mathbb{C} \times \mathbb{C}^d \times \mathbb{C}^d$,

$$F v_i = \lambda_i v_i, \; u_i^* F = \lambda_i u_i^* \; \text{ and } \; u_i^* v_j = \delta_{ij} \quad \Longrightarrow \quad F = \sum_{i\in[d]} \lambda_i v_i u_i^*$$

- Expected dynamics: $\mathbb{E}[X_t \,|\, X_0 = x] = F^t x = \sum_{i\in[d]} \lambda_i^t \, (u_i^* x) \, v_i$

A different perspective via measurements/observables :

- $\mathcal{F} := \{ f_w := \langle \cdot, w \rangle : \mathbb{R}^d \to \mathbb{R} \,|\, w \in \mathbb{R}^d \} \quad \Longrightarrow \quad \mathbb{E}[f_w(X_{t+1}) \,|\, X_t = x] = \langle x, F^* w \rangle = f_{F^* w}(x)$

- Expected dynamics of observables :

$$f_w \in \mathcal{F} \quad \Longrightarrow \quad \mathbb{E}[f_w(X_t) \,|\, X_0 = x] = \langle x, (F^*)^t w \rangle = \sum_{i\in[d]} \lambda_i^t \, \langle f_w, f_{v_i} \rangle_{\mathcal{F}} \, f_{u_i}(x)$$

- Even when F is not linear, the mapping $f \mapsto \mathbb{E}[f(X_{t+1}) \,|\, X_t = x]$ is linear!

Koopman Operator Framework

- Let $\{X_t : t \in \mathbb{N}\}$ be a time-homogeneous Markov chain,

$$\mathbb{P}\{X_{t+1} \in B \mid X_t = x\} = \underbrace{p(x,\ B)}_{\text{transition kernel}}\ , \qquad (x, B) \in \mathcal{X} \times \Sigma_{\mathcal{X}},\ t \in \mathbb{N}$$

Koopman Operator Framework

- Let $\{X_t \colon t \in \mathbb{N}\}$ be a time-homogeneous Markov chain,

$$\mathbb{P}\{X_{t+1} \in B \mid X_t = x\} = \underbrace{p(x,\ B)}_{\text{transition kernel}}\ , \quad (x, B) \in \mathcal{X} \times \Sigma_{\mathcal{X}},\ t \in \mathbb{N}$$

- If $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ is a vector space of observables its Koopman operator $A_{\mathcal{F}} \colon \mathcal{F} \to \mathcal{F}$ is

$$[A_{\mathcal{F}} f](x) := \int_{\mathcal{X}} p(x, dy) f(y) = \mathbb{E}\left[f(X_{t+1}) | X_t = x\right], \quad f \in \mathcal{F},\ x \in \mathcal{X}$$

Koopman Operator Framework

- Let $\{X_t\colon t\in\mathbb{N}\}$ be a time-homogeneous Markov chain,

$$\mathbb{P}\{X_{t+1}\in B \mid X_t = x\} = \underbrace{p(x,\ B)}_{\text{transition kernel}}\ , \qquad (x,B)\in\mathcal{X}\times\Sigma_{\mathcal{X}},\ t\in\mathbb{N}$$

- If $\mathcal{F}\subset\mathbb{R}^{\mathcal{X}}$ is a vector space of observables its Koopman operator $A_{\mathcal{F}}\colon\mathcal{F}\to\mathcal{F}$ is

$$[A_{\mathcal{F}}f](x) := \int_{\mathcal{X}} p(x,dy)f(y) = \mathbb{E}\left[f(X_{t+1})|X_t = x\right], \quad f\in\mathcal{F},\, x\in\mathcal{X}$$

- We can use spectral theory: if $\exists(\mu_i,g_i,f_i)\in\mathbb{C}\times\mathcal{F}\times\mathcal{F}$, $i\in\mathbb{N}$, s.t.

$$A_{\mathcal{F}}f_i = \mu_i f_i, \qquad A_{\mathcal{F}}^{*}g_i = \bar{\mu}_i g_i, \qquad \langle f_i,\bar{g}_j\rangle = \delta_{ij}, \quad i,j\in\mathbb{N}$$

  then the Koopman Mode Decomposition of $f\in\operatorname{span}\{f_1,f_2,\ldots\}$:

$$[A_{\mathcal{F}}^{t}f](x) = \mathbb{E}[f(X_t)\,|\,X_0 = x] = \sum_i \mu_i^t\,\langle f,\bar{g}_i\rangle\, f_i(x), \quad x\in\mathcal{X},\, t\in\mathbb{N}$$

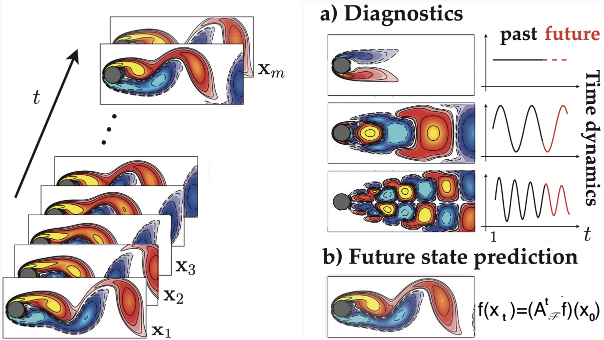
Koopman Mode Decomposition (KMD)

$$[A^t_{\mathcal{F}} f](x) = \mathbb{E}[f(X_t) \,|\, X_0 = x] = \sum_i \mu_i^t \, \langle f, \bar{g}_i \rangle f_i(x), \quad x \in \mathcal{X},\, t \in \mathbb{N}$$

- Time oscillations λ_i^t with amplitudes $|\lambda_i|^t$ and frequencies $e^{i \mathrm{Arg}(\lambda_i) t}$, *i.e.*

$$\frac{\Im(\ln \lambda_i)}{2\pi \Delta t}$$

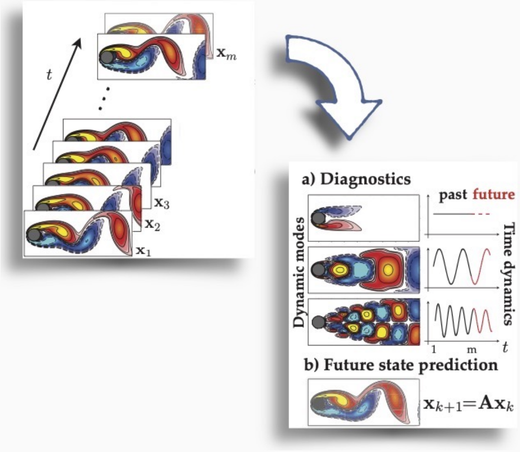
- Static modes $\langle f, \bar{\xi}_i \rangle$ of observable f
- Terms $\psi_i(x)$ depending only on the initial condition



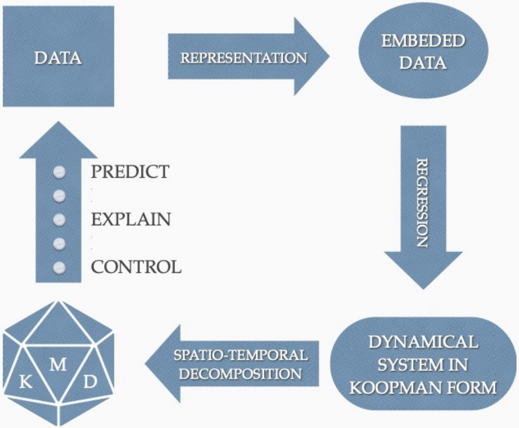
(Picture from [Kutz et al. 2016])

Our Approach

Let's use the **kernel trick** - replace $\langle x, y \rangle$ with $k(x, y) = \langle \phi(x), \phi(y) \rangle$!

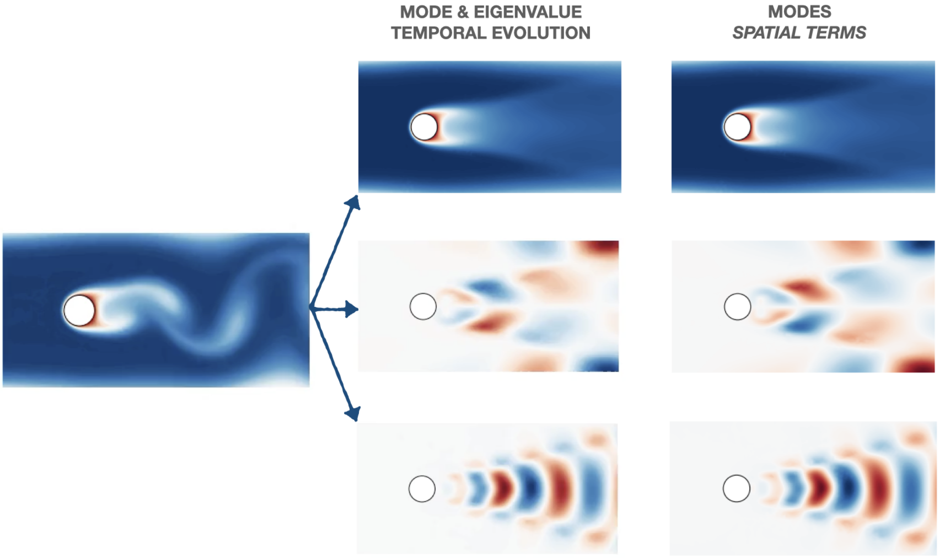


(Picture from [Kutz et al. 2016])



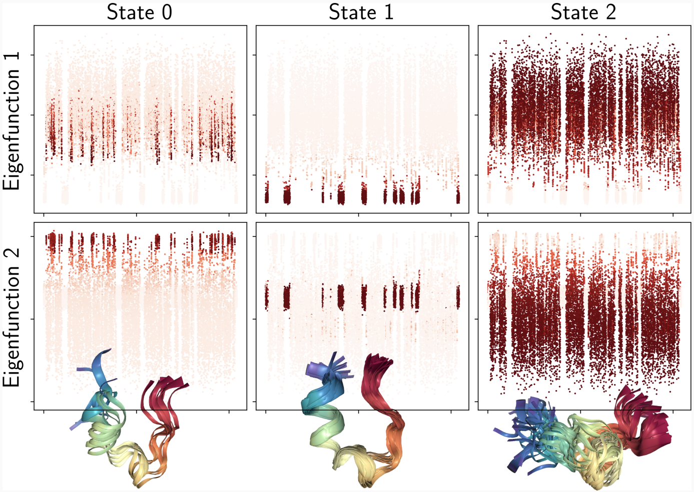
KOR GitHub page **kooplearn** SciKit Learn compliant & KeOps implementations

kooplearn in action...



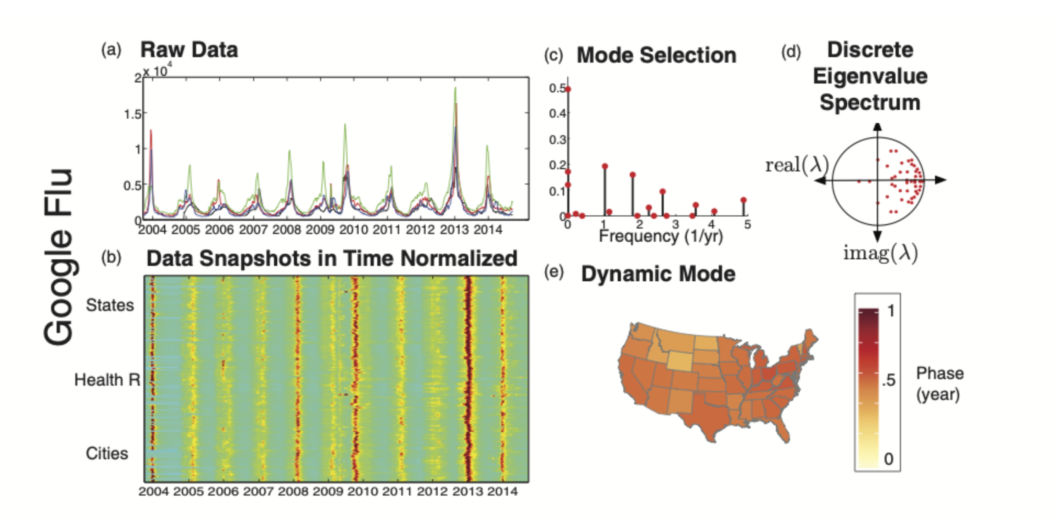
Some interesting applications

Molecular Dynamics



(Picture from [Meanti et al. 2023])

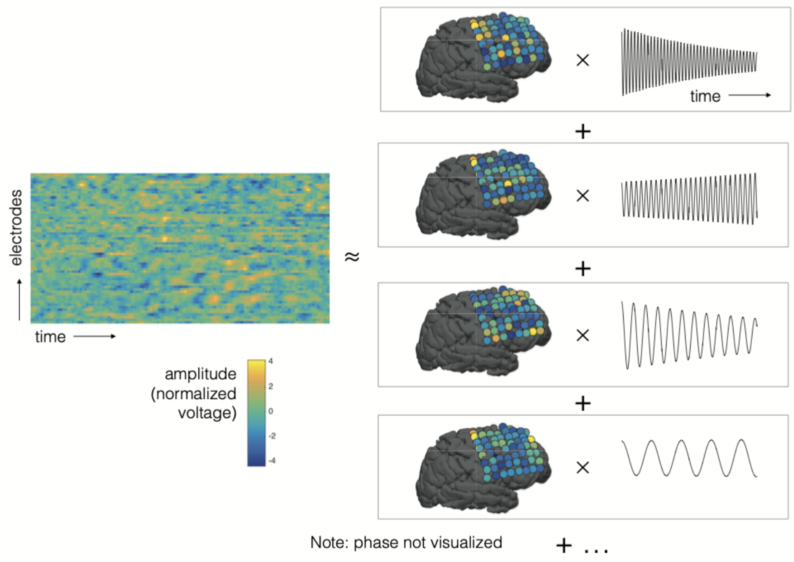
Epidemiology



(Picture from [Kutz et al. 2016])

Koopman modes give insights into spatio-temporal correlations

Neuroscience



(Picture from [Kutz et al. 2016])

Related Work (list by far incomplete!)

Data-driven algorithms to reconstruct dynamical systems:

- Williams, Rowley, Kevrekidis (2015). A kernel-based method for data-driven Koopman spectral analysis. *J. of Computational Dynamics*
- Kutz, Brunton, Brunton, Proctor (2016). *Dynamic Mode Decomposition*. SIAM.
- Klus, Schuster and Muandet (2019) Eigendecompositions of transfer operators in reproducing kernel Hilbert spaces. *Journal of Nonlinear Science*

Koopman operator theory:

- Brunton, Budišić, Kaiser, Kutz (2022). Modern Koopman Theory for Dynamical Systems. *SIAM Review*
- Budišić, Mohr, Mezić (2012). Applied Koopmanism. *Chaos: An Interdisciplinary J. of Nonlinear Science*
- Das and Giannakis (2020). Koopman spectra in reproducing kernel Hilbert spaces. *Applied and Computational Harmonic Analysis*

Statistical learning / link to CME (see below):

- Grünewälder *et al.* (2012). Conditional mean embeddings as regressors. *ICML*
- Muandet, Fukumizu, Sriperumbudur and Schölkopf (2017). Kernel Mean Embedding of Distributions: A Review and Beyond. *Foundations and Trends in Machine Learning*
- Li, Meunier, Mollenhauer and Gretton (2022). Optimal rates for regularized conditional mean embedding learning. *NeurIPS*

Statistical Learning Framework

Which Koopman are We Learning ?

$$A_{\mathcal{F}} : \mathcal{F} \to \mathcal{F}, \qquad (A_{\mathcal{F}} f)(x) = \int_{\mathcal{X}} p(x, dy) f(y) = \mathbb{E}\left[f(X_{t+1}) | X_t = x\right]$$

- **What is an appropriate $\mathcal{F}$?**

Which Koopman are We Learning ?

$$A_{\mathcal{F}} \colon \mathcal{F} \to \mathcal{F}, \qquad (A_{\mathcal{F}} f)(x) = \int_{\mathcal{X}} p(x, dy) f(y) = \mathbb{E}\left[f(X_{t+1}) | X_t = x\right]$$

- **What is an appropriate $\mathcal{F}$?** Assuming invariant distribution π:

$$\pi(B) = \int_{\mathcal{X}} \underbrace{\pi(dx)\, p(x, B)}_{\text{joint distribution } \rho}, \quad \forall\, B \in \Sigma_{\mathcal{X}}$$

we can choose $\mathcal{F} = L^2_\pi(\mathcal{X})$, and denote $A_\pi \equiv A_{L^2_\pi(\mathcal{X})}$. In general $\|A_\pi\| = 1$ and $A_\pi f = f$, for π-a.e. constant function f!

Which Koopman are We Learning ?

$$A_{\mathcal{F}} \colon \mathcal{F} \to \mathcal{F}, \qquad (A_{\mathcal{F}} f)(x) = \int_{\mathcal{X}} p(x, dy) f(y) = \mathbb{E}\left[f(X_{t+1}) | X_t = x\right]$$

- **What is an appropriate $\mathcal{F}$?** Assuming invariant distribution π:

$$\pi(B) = \int_{\mathcal{X}} \underbrace{\pi(dx)\, p(x, B)}_{\text{joint distribution } \rho}, \quad \forall\, B \in \Sigma_{\mathcal{X}}$$

we can choose $\mathcal{F} = L^2_\pi(\mathcal{X})$, and denote $A_\pi \equiv A_{L^2_\pi(\mathcal{X})}$. In general $\|A_\pi\| = 1$ and $A_\pi f = f$, for π-a.e. constant function f!

Our example: If $F = F^*$ and $\|F\| < 1$, then $\pi \equiv \mathcal{N}(0, C)$ for $C = \sigma^2 (I - F^2)^{-1}$

Which Koopman are We Learning ?

$$A_{\mathcal{F}} : \mathcal{F} \to \mathcal{F}, \qquad (A_{\mathcal{F}} f)(x) = \int_{\mathcal{X}} p(x, dy) f(y) = \mathbb{E}\left[f(X_{t+1}) | X_t = x\right]$$

- **What is an appropriate $\mathcal{F}$?** Assuming invariant distribution π:

$$\pi(B) = \int_{\mathcal{X}} \underbrace{\pi(dx)\, p(x, B)}_{\text{joint distribution } \rho}, \quad \forall\, B \in \Sigma_{\mathcal{X}}$$

  we can choose $\mathcal{F} = L^2_\pi(\mathcal{X})$, and denote $A_\pi \equiv A_{L^2_\pi(\mathcal{X})}$. In general $\|A_\pi\| = 1$ and $A_\pi f = f$, for π-a.e. constant function f!

  Our example: If $F = F^*$ and $\|F\| < 1$, then $\pi \equiv \mathcal{N}(0, C)$ for $C = \sigma^2(I - F^2)^{-1}$

- **How to learn A_π from data when not even a domain is available?**

Which Koopman are We Learning ?

$$A_{\mathcal{F}}\colon \mathcal{F} \to \mathcal{F}, \qquad (A_{\mathcal{F}} f)(x) = \int_{\mathcal{X}} p(x, dy) f(y) = \mathbb{E}\left[f(X_{t+1}) | X_t = x\right]$$

- **What is an appropriate $\mathcal{F}$?** Assuming invariant distribution π:

$$\pi(B) = \int_{\mathcal{X}} \underbrace{\pi(dx)\, p(x, B)}_{\text{joint distribution } \rho}, \quad \forall\, B \in \Sigma_{\mathcal{X}}$$

  we can choose $\mathcal{F} = L^2_\pi(\mathcal{X})$, and denote $A_\pi \equiv A_{L^2_\pi(\mathcal{X})}$. In general $\|A_\pi\| = 1$ and $A_\pi f = f$, for π-a.e. constant function f!

  Our example: If $F = F^*$ and $\|F\| < 1$, then $\pi \equiv \mathcal{N}(0, C)$ for $C = \sigma^2 (I - F^2)^{-1}$

- **How to learn A_π from data when not even a domain is available?**
  - $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive definite kernel such that $k(\cdot, \cdot) < \infty$ π-a.e.

Which Koopman are We Learning ?

$$A_{\mathcal{F}} : \mathcal{F} \to \mathcal{F}, \qquad (A_{\mathcal{F}} f)(x) = \int_{\mathcal{X}} p(x, dy) f(y) = \mathbb{E}\left[f(X_{t+1}) | X_t = x\right]$$

- **What is an appropriate $\mathcal{F}$?** Assuming invariant distribution π:

$$\pi(B) = \int_{\mathcal{X}} \underbrace{\pi(dx)\, p(x, B)}_{\text{joint distribution } \rho}, \quad \forall\, B \in \Sigma_{\mathcal{X}}$$

  we can choose $\mathcal{F} = L^2_\pi(\mathcal{X})$, and denote $A_\pi \equiv A_{L^2_\pi(\mathcal{X})}$. In general $\|A_\pi\| = 1$ and $A_\pi f = f$, for π-a.e. constant function f!

  Our example: If $F = F^*$ and $\|F\| < 1$, then $\pi \equiv \mathcal{N}(0, C)$ for $C = \sigma^2 (I - F^2)^{-1}$

- **How to learn A_π from data when not even a domain is available?**
  - $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive definite kernel such that $k(\cdot, \cdot) < \infty$ π-a.e.
  - $\mathcal{H}$ the associated reproducing kernel Hilbert spaces (RKHS) is then $\mathcal{H} \subseteq L^2_\pi(\mathcal{X})$

Which Koopman are We Learning ?

$$A_{\mathcal{F}} \colon \mathcal{F} \to \mathcal{F}, \qquad (A_{\mathcal{F}} f)(x) = \int_{\mathcal{X}} p(x, dy) f(y) = \mathbb{E}\left[f(X_{t+1}) | X_t = x\right]$$

- **What is an appropriate $\mathcal{F}$?** Assuming invariant distribution π:

$$\pi(B) = \int_{\mathcal{X}} \underbrace{\pi(dx)\, p(x, B)}_{\text{joint distribution } \rho}, \quad \forall\, B \in \Sigma_{\mathcal{X}}$$

we can choose $\mathcal{F} = L^2_\pi(\mathcal{X})$, and denote $A_\pi \equiv A_{L^2_\pi(\mathcal{X})}$. In general $\|A_\pi\| = 1$ and $A_\pi f = f$, for π-a.e. constant function f!

Our example: If $F = F^*$ and $\|F\| < 1$, then $\pi \equiv \mathcal{N}(0, C)$ for $C = \sigma^2(I - F^2)^{-1}$

- **How to learn A_π from data when not even a domain is available?**
    - $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive definite kernel such that $k(\cdot, \cdot) < \infty$ π-a.e.
    - $\mathcal{H}$ the associated reproducing kernel Hilbert spaces (RKHS) is then $\mathcal{H} \subseteq L^2_\pi(\mathcal{X})$
    - with feature map $\phi(x) := k(x, \cdot)$ we form subspaces from data $(x_i)_i$ by $\sum_i c_i \phi(x_i)$

Which Koopman are We Learning ?

$$A_{\mathcal{F}} : \mathcal{F} \to \mathcal{F}, \qquad (A_{\mathcal{F}} f)(x) = \int_{\mathcal{X}} p(x, dy) f(y) = \mathbb{E}\left[f(X_{t+1}) | X_t = x\right]$$

- **What is an appropriate $\mathcal{F}$?** Assuming invariant distribution π:

$$\pi(B) = \int_{\mathcal{X}} \underbrace{\pi(dx)\, p(x, B)}_{\text{joint distribution } \rho}, \quad \forall\, B \in \Sigma_{\mathcal{X}}$$

  we can choose $\mathcal{F} = L^2_\pi(\mathcal{X})$, and denote $A_\pi \equiv A_{L^2_\pi(\mathcal{X})}$. In general $\|A_\pi\| = 1$ and $A_\pi f = f$, for π-a.e. constant function f!

  Our example: If $F = F^*$ and $\|F\| < 1$, then $\pi \equiv \mathcal{N}(0, C)$ for $C = \sigma^2(I - F^2)^{-1}$

- **How to learn A_π from data when not even a domain is available?**
  - $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive definite kernel such that $k(\cdot, \cdot) < \infty$ π-a.e.
  - $\mathcal{H}$ the associated reproducing kernel Hilbert spaces (RKHS) is then $\mathcal{H} \subseteq L^2_\pi(\mathcal{X})$
  - with feature map $\phi(x) := k(x, \cdot)$ we form subspaces from data $(x_i)_i$ by $\sum_i c_i \phi(x_i)$
  - we use the reproducing property $h(x) = \langle \phi(x), h \rangle_{\mathcal{H}}$, also known as a "kernel trick"

Statistical Learning Framework

- **Let's start with a notion of risk of a potential estimator $G\colon \mathcal{H} \to \mathcal{H}$:**

$$\mathcal{R}(G) = \mathbb{E}\Big[\sum_{i\in\mathbb{N}} (h_i(X_{t+1}) - (Gh_i)(X_t))^2\Big] \text{ i.e.}$$

the cumulative expected one-step-ahead prediction error over an o.n. basis $(h_i)_{i\in\mathbb{N}}$ of $\mathcal{H}$.

Statistical Learning Framework

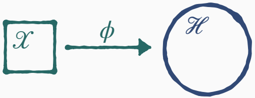
- **Let's start with a notion of risk of a potential estimator $G\colon \mathcal{H} \to \mathcal{H}$:**

$$\mathcal{R}(G) = \mathbb{E}\Big[\sum_{i\in\mathbb{N}} (h_i(X_{t+1}) - (Gh_i)(X_t))^2\Big] \text{ i.e.}$$

  the cumulative expected one-step-ahead prediction error over an o.n. basis $(h_i)_{i\in\mathbb{N}}$ of $\mathcal{H}$.

- **Kernel trick:** Embed data and aim to learn $G\colon \mathcal{H} \to \mathcal{H}$ s.t.

$$G^*\phi(X) \approx \phi(Y), \quad (X,Y) \sim \rho$$



Statistical Learning Framework

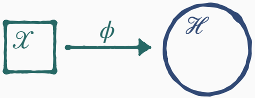
- **Let's start with a notion of risk of a potential estimator** $G\colon \mathcal{H} \to \mathcal{H}$**:**

$$\mathcal{R}(G) = \mathbb{E}\Big[\sum_{i\in\mathbb{N}} (h_i(X_{t+1}) - (Gh_i)(X_t))^2\Big] \quad \text{i.e.}$$

  the cumulative expected one-step-ahead prediction error over an o.n. basis $(h_i)_{i\in\mathbb{N}}$ of $\mathcal{H}$.

- **Kernel trick:** Embed data and aim to learn $G\colon \mathcal{H} \to \mathcal{H}$ s.t.

$$G^*\phi(X) \approx \phi(Y), \quad (X, Y) \sim \rho$$



- The **risk has equivalent form** $\mathcal{R}(G) := \mathbb{E}_{(X,Y)\sim\rho}\|\phi(Y) - G^*\phi(X)\|^2$

Statistical Learning Framework

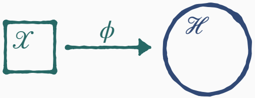
- **Let's start with a notion of risk of a potential estimator $G\colon \mathcal{H} \to \mathcal{H}$:**

$$\mathcal{R}(G) = \mathbb{E}\Big[\sum_{i\in\mathbb{N}} (h_i(X_{t+1}) - (Gh_i)(X_t))^2\Big] \text{ i.e.}$$

the cumulative expected one-step-ahead prediction error over an o.n. basis $(h_i)_{i\in\mathbb{N}}$ of $\mathcal{H}$.

- **Kernel trick:** Embed data and aim to learn $G\colon \mathcal{H} \to \mathcal{H}$ s.t.

$$G^*\phi(X) \approx \underbrace{\mathbb{E}[\phi(X_{t+1}) \,|\, X_t = X]}_{g_p(X)}, \quad X \sim \pi$$



- The **risk has equivalent form** $\mathcal{R}(G) := \mathbb{E}_{(X,Y)\sim\rho}\|\phi(Y) - G^*\phi(X)\|^2$

- and we have the **bias-variance** decomposition

$$\underbrace{\mathbb{E}_{(X,Y)\sim\rho}\|\phi(Y) - G^*\phi(X)\|^2}_{\mathcal{R}(G)} = \underbrace{\mathbb{E}_{X\sim\pi}\|g_p(X) - G^*\phi(X)\|^2}_{\text{excess risk}} + \underbrace{\mathbb{E}_{(X,Y)\sim\rho}\|g_p(X) - \phi(Y)\|^2}_{\text{irreducible risk } \mathcal{R}_0}$$

Statistical Learning Framework

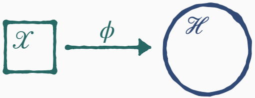
- **Let's start with a notion of risk of a potential estimator** $G\colon \mathcal{H} \to \mathcal{H}$:

$$\mathcal{R}(G) = \mathbb{E}\Big[\sum_{i\in\mathbb{N}} (h_i(X_{t+1}) - (Gh_i)(X_t))^2\Big] \text{ i.e.}$$

the cumulative expected one-step-ahead prediction error over an o.n. basis $(h_i)_{i\in\mathbb{N}}$ of $\mathcal{H}$.

- **Kernel trick:** Embed data and aim to learn $G\colon \mathcal{H} \to \mathcal{H}$ s.t.

$$G^*\phi(X) \approx \underbrace{\mathbb{E}[\phi(X_{t+1})\,|\,X_t = X]}_{g_p(X)}, \quad X \sim \pi$$



- The **risk has equivalent form** $\mathcal{R}(G) := \mathbb{E}_{(X,Y)\sim\rho}\|\phi(Y) - G^*\phi(X)\|^2$

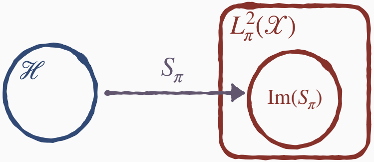
- and we have the **bias-variance** decomposition

$$\underbrace{\mathbb{E}_{(X,Y)\sim\rho}\|\phi(Y) - G^*\phi(X)\|^2}_{\mathcal{R}(G)} = \underbrace{\mathbb{E}_{X\sim\pi}\|g_p(X) - G^*\phi(X)\|^2}_{\text{excess risk}} + \underbrace{\mathbb{E}_{(X,Y)\sim\rho}\|g_p(X) - \phi(Y)\|^2}_{\text{irreducible risk } \mathcal{R}_0}$$

- g_p is known as the **conditional mean embedding (CME)** of transition kernel p into $\mathcal{H}$!

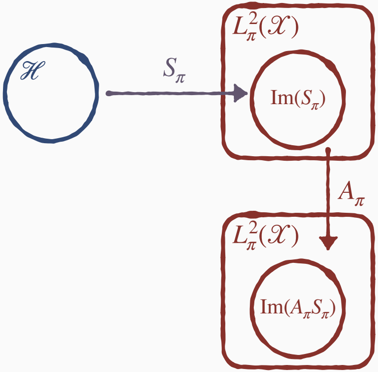
Statistical Learning Framework

- Since $k(\cdot,\cdot) \in L^2_\pi(\mathcal{X})$ then $\mathcal{H} \subseteq L^2_\pi(\mathcal{X})$, so the injection operator S_π is Hilbert-Schmidt



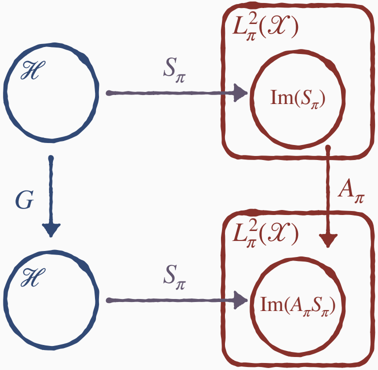
Statistical Learning Framework

- Since $k(\cdot,\cdot) \in L^2_\pi(\mathcal{X})$ then $\mathcal{H} \subseteq L^2_\pi(\mathcal{X})$, so the injection operator S_π is Hilbert-Schmidt
- The restriction of the Koopman operator to $\mathcal{H}$ $A_{\pi\,|_\mathcal{H}} \equiv A_\pi S_\pi$ is then Hilbert-Schmidt, too!



Statistical Learning Framework

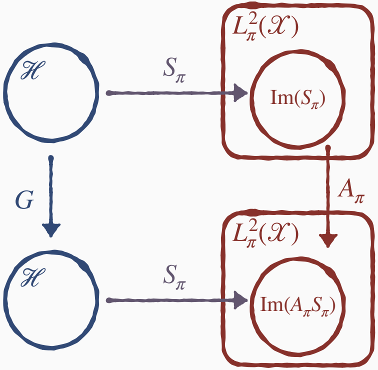
- Since $k(\cdot, \cdot) \in L^2_\pi(\mathcal{X})$ then $\mathcal{H} \subseteq L^2_\pi(\mathcal{X})$, so the injection operator S_π is Hilbert-Schmidt
- The restriction of the Koopman operator to $\mathcal{H}$ $A_{\pi|_\mathcal{H}} \equiv A_\pi S_\pi$ is then Hilbert-Schmidt, too!
- We should solve the inverse problem $S_\pi G = A_\pi S_\pi$



Statistical Learning Framework

- Since $k(\cdot,\cdot) \in L^2_\pi(\mathcal{X})$ then $\mathcal{H} \subseteq L^2_\pi(\mathcal{X})$, so the injection operator S_π is Hilbert-Schmidt
- The restriction of the Koopman operator to $\mathcal{H}$ $A_{\pi\,|_{\mathcal{H}}} \equiv A_\pi S_\pi$ is then Hilbert-Schmidt, too!
- We should solve the inverse problem $S_\pi G = A_\pi S_\pi$
- But, since the risk can be decomposed as

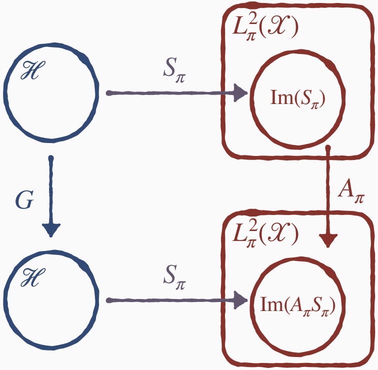
$$\mathcal{R}(G) = \underbrace{\|A_\pi S_\pi - S_\pi G\|^2_{\mathrm{HS}}}_{\mathcal{E}^2_{\mathrm{HS}}(G)} + \underbrace{\|S_\pi\|^2_{\mathrm{HS}} - \|A_\pi S_\pi\|^2_{\mathrm{HS}}}_{\mathcal{R}_0}$$



Statistical Learning Framework

- Since $k(\cdot,\cdot) \in L^2_\pi(\mathcal{X})$ then $\mathcal{H} \subseteq L^2_\pi(\mathcal{X})$, so the injection operator S_π is Hilbert-Schmidt
- The restriction of the Koopman operator to $\mathcal{H}$ $A_{\pi\,|_{\mathcal{H}}} \equiv A_\pi S_\pi$ is then Hilbert-Schmidt, too!
- We should solve the inverse problem $S_\pi G = A_\pi S_\pi$
- But, since the risk can be decomposed as

$$\mathcal{R}(G) = \underbrace{\|A_\pi S_\pi - S_\pi G\|^2_{\mathrm{HS}}}_{\mathcal{E}^2_{\mathrm{HS}}(G)} + \underbrace{\|S_\pi\|^2_{\mathrm{HS}} - \|A_\pi S_\pi\|^2_{\mathrm{HS}}}_{\mathcal{R}_0}$$



the problem of learning $A_{\pi\,|_{\mathcal{H}}}$ by $S_\pi G$ is equivalent to learning g_p by $G^*\phi(\cdot)$!
Duality with CME is via reproducing property $[A_{\pi\,|_{\mathcal{H}}} h](x) = \langle h, g_p(x)\rangle_{\mathcal{H}}$

Statistical Learning Framework

- Since $k(\cdot,\cdot) \in L^2_\pi(\mathcal{X})$ then $\mathcal{H} \subseteq L^2_\pi(\mathcal{X})$, so the injection operator S_π is Hilbert-Schmidt

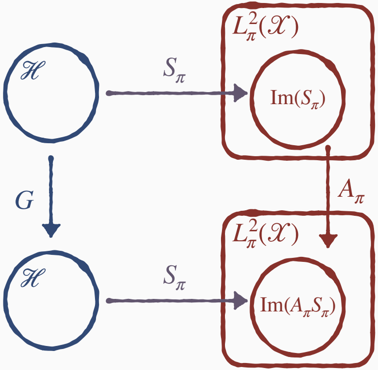
- The restriction of the Koopman operator to $\mathcal{H}$ $A_{\pi\,|\mathcal{H}} \equiv A_\pi S_\pi$ is then Hilbert-Schmidt, too!

- We should solve the inverse problem $S_\pi G = A_\pi S_\pi$

- But, since the risk can be decomposed as

$$\mathcal{R}(G) = \underbrace{\|A_\pi S_\pi - S_\pi G\|^2_{\mathrm{HS}}}_{\mathcal{E}^2_{\mathrm{HS}}(G)} + \underbrace{\|S_\pi\|^2_{\mathrm{HS}} - \|A_\pi S_\pi\|^2_{\mathrm{HS}}}_{\mathcal{R}_0}$$

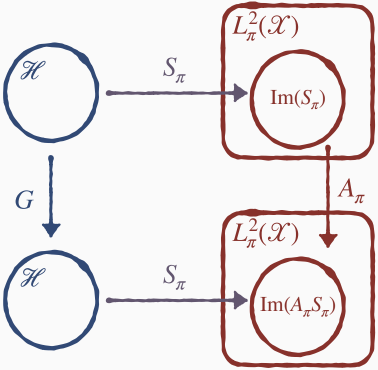
the problem of learning $A_{\pi\,|\mathcal{H}}$ by $S_\pi G$ is equivalent to learning g_p by $G^*\phi(\cdot)$!
Duality with CME is via reproducing property $[A_{\pi\,|\mathcal{H}} h](x) = \langle h, g_p(x)\rangle_{\mathcal{H}}$



- **How well can we learn A_π via $\mathcal{H}$?**

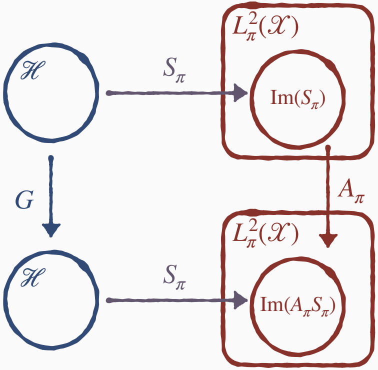
Statistical Learning Framework

- **Proposition:** If $P_{\mathcal{H}}$ is orthogonal projector onto $\mathrm{cl}(\mathrm{Im}(S_\pi))$ in $L^2_\pi(\mathcal{X})$, then for every $\delta > 0$ there exists a finite rank non-defective operator G such that $\|A_\pi S_\pi - S_\pi G\|^2_{\mathrm{HS}} \leq \|[I - P_{\mathcal{H}}]A_\pi S_\pi\|^2_{\mathrm{HS}} + \delta$.



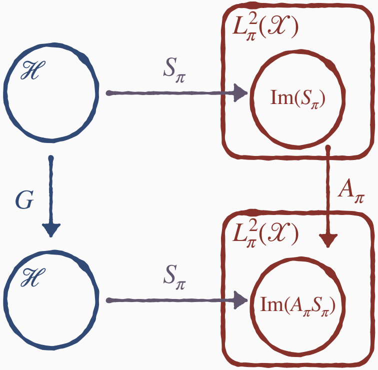
Statistical Learning Framework

- **Proposition:** If $P_{\mathcal{H}}$ is orthogonal projector onto $\mathrm{cl}(\mathrm{Im}(S_\pi))$ in $L^2_\pi(\mathcal{X})$, then for every $\delta > 0$ there exists a finite rank non-defective operator G such that $\|A_\pi S_\pi - S_\pi G\|^2_{\mathrm{HS}} \leq \|[I - P_{\mathcal{H}}]A_\pi S_\pi\|^2_{\mathrm{HS}} + \delta$.
- **Remark:** $\mathrm{Im}(A_\pi S_\pi) \subseteq \mathrm{cl}(\mathrm{Im}(S_\pi))$ holds for $\mathcal{H}$ that is dense in $L^2_\pi(\mathcal{X})$ (i.e. for *universal k*) which implies that $P_{\mathcal{H}} = I$.



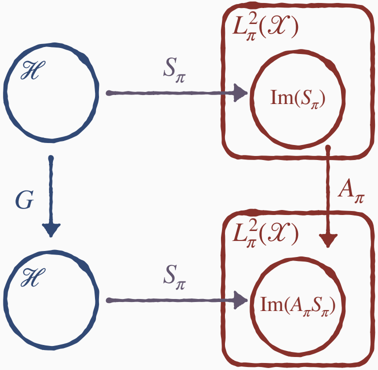
Statistical Learning Framework

- **Proposition:** If $P_{\mathcal{H}}$ is orthogonal projector onto $\mathrm{cl}(\mathrm{Im}(S_\pi))$ in $L^2_\pi(\mathcal{X})$, then for every $\delta > 0$ there exists a finite rank non-defective operator G such that $\|A_\pi S_\pi - S_\pi G\|^2_{\mathrm{HS}} \leq \|[I - P_{\mathcal{H}}]A_\pi S_\pi\|^2_{\mathrm{HS}} + \delta$.
- **Remark:** $\mathrm{Im}(A_\pi S_\pi) \subseteq \mathrm{cl}(\mathrm{Im}(S_\pi))$ holds for $\mathcal{H}$ that is dense in $L^2_\pi(\mathcal{X})$ (i.e. for *universal* k) which implies that $P_{\mathcal{H}} = I$.
- Two cases arise depending on whether $\inf_{G \in \mathrm{HS}(\mathcal{H})} \mathcal{E}(G)$ is attained or not:



Statistical Learning Framework

- **Proposition:** If $P_{\mathcal{H}}$ is orthogonal projector onto $\mathrm{cl}(\mathrm{Im}(S_\pi))$ in $L^2_\pi(\mathcal{X})$, then for every $\delta > 0$ there exists a finite rank non-defective operator G such that $\|A_\pi S_\pi - S_\pi G\|^2_{\mathrm{HS}} \leq \|[I - P_{\mathcal{H}}]A_\pi S_\pi\|^2_{\mathrm{HS}} + \delta$.
- **Remark:** $\mathrm{Im}(A_\pi S_\pi) \subseteq \mathrm{cl}(\mathrm{Im}(S_\pi))$ holds for $\mathcal{H}$ that is dense in $L^2_\pi(\mathcal{X})$ (i.e. for *universal* k) which implies that $P_{\mathcal{H}} = I$.
- Two cases arise depending on whether $\inf_{G\in\mathrm{HS}(\mathcal{H})} \mathcal{E}(G)$ is attained or not:

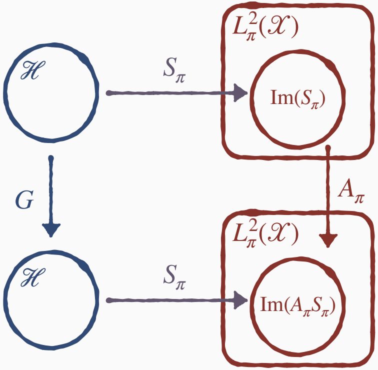


(i) well-specified case, there exists π-a.e. Koopman operator $G_{\mathcal{H}} := C^\dagger T \in \mathrm{HS}(\mathcal{H})$, where $C := \mathbb{E}_{X\sim\pi}\phi(X)\otimes\phi(X)$ and $T := \mathbb{E}_{(X,Y)\sim\rho}\phi(X)\otimes\phi(Y)$, i.e.

$$A_\pi S_\pi = S_\pi G_{\mathcal{H}} \iff G_{\mathcal{H}} f = \mathbb{E}[f(X_{t+1}) \,|\, X_t = \cdot] \quad \pi\text{-a.e. for every } f \in \mathcal{H}.$$

Statistical Learning Framework

- **Proposition:** If $P_{\mathcal{H}}$ is orthogonal projector onto $\mathrm{cl}(\mathrm{Im}(S_\pi))$ in $L^2_\pi(\mathcal{X})$, then for every $\delta > 0$ there exists a finite rank non-defective operator G such that $\|A_\pi S_\pi - S_\pi G\|^2_{\mathrm{HS}} \leq \|[I - P_{\mathcal{H}}]A_\pi S_\pi\|^2_{\mathrm{HS}} + \delta$.
- **Remark:** $\mathrm{Im}(A_\pi S_\pi) \subseteq \mathrm{cl}(\mathrm{Im}(S_\pi))$ holds for $\mathcal{H}$ that is dense in $L^2_\pi(\mathcal{X})$ (i.e. for *universal* k) which implies that $P_{\mathcal{H}} = I$.
- Two cases arise depending on whether $\inf_{G \in \mathrm{HS}(\mathcal{H})} \mathcal{E}(G)$ is attained or not:



(i) well-specified case, there exists π-a.e. Koopman operator $G_{\mathcal{H}} := C^\dagger T \in \mathrm{HS}\,(\mathcal{H})$, where $C := \mathbb{E}_{X\sim\pi}\phi(X) \otimes \phi(X)$ and $T := \mathbb{E}_{(X,Y)\sim\rho}\phi(X) \otimes \phi(Y)$, i.e.

$$A_\pi S_\pi = S_\pi G_{\mathcal{H}} \quad \Longleftrightarrow \quad G_{\mathcal{H}} f = \mathbb{E}[f(X_{t+1}) \,|\, X_t = \cdot\,] \quad \pi\text{-a.e. for every } f \in \mathcal{H}.$$

(ii) misspecified case, $\mathcal{H}$ does not admit a HS π-a.e. Koopman operator $\mathcal{H} \to \mathcal{H}$

Empirical Estimators and Statistical Bounds

Empirical Estimators of A_π

- We either observe an i.i.d. $\mathcal{D} = (x_i, y_i)_{i=1}^n$ from ρ, or from a trajectory $\ldots, x_i, \underbrace{x_{i+1}}_{y_i}, \ldots$

Empirical Estimators of A_π

- We either observe an i.i.d. $\mathcal{D} = (x_i, y_i)_{i=1}^n$ from ρ, or from a trajectory $\ldots, x_i, \underbrace{x_{i+1}}_{y_i}, \ldots$
- The Koopman Operator Regression is then: Given the data $\mathcal{D}$ solve $\min_{G \in \mathrm{HS}(\mathcal{H})} \mathcal{R}(G)$

Empirical Estimators of A_π

- We either observe an i.i.d. $\mathcal{D} = (x_i, y_i)_{i=1}^n$ from ρ, or from a trajectory $\ldots, x_i, \underbrace{x_{i+1}}_{y_i}, \ldots$
- The Koopman Operator Regression is then: Given the data $\mathcal{D}$ solve $\min_{G \in \mathrm{HS}(\mathcal{H})} \mathcal{R}(G)$
- Different estimators arise by minimizing over a set of operators the empirical risk

$$\widehat{\mathcal{R}}(G) := \frac{1}{n} \sum_{i=1}^{n} \|\phi(y_i) - G^* \phi(x_i)\|_{\mathcal{H}}^2$$

or, equivalently,

$$\widehat{\mathcal{R}}(G) \equiv \|\widehat{Z} - \widehat{S} G\|_{\mathrm{HS}}^2$$

using the sampling operators $\widehat{S}, \widehat{Z} \in \mathrm{HS}\left(\mathcal{H}, \mathbb{R}^n\right)$ of inputs and outputs

$$\widehat{S} f = \left(n^{-\frac{1}{2}} f(x_i)\right)_{i=1}^{n}, \qquad \widehat{Z} f = \left(n^{-\frac{1}{2}} f(y_i)\right)_{i=1}^{n}$$

that lead to covariance and cross-covariance operators

$$\widehat{C} = \widehat{S}^* \widehat{S} = \frac{1}{n} \sum_{i \in [n]} \phi(x_i) \otimes \phi(x_i) \quad \text{and } \widehat{T} = \widehat{S}^* \widehat{Z} = \frac{1}{n} \sum_{i \in [n]} \phi(x_i) \otimes \phi(y_i)$$

Estimators via ERM

The estimators have the form $\widehat{G} = \widehat{S}^* W \widehat{Z}, \quad W \in \mathbb{R}^{n \times n}$

$$\min_{\widehat{G} \in \mathrm{HS}(\mathcal{H})} \widehat{\mathcal{R}}(\widehat{G}) + \gamma \|\widehat{G}\|_{\mathrm{HS}}^2$$

- Kernel Ridge Regression (KRR) $G_\gamma := C_\gamma^{-1} T$:
  $W = K_\gamma^{-1}$, with $K = (k(x_i, x_j))_{i,j=1}^n$, $K_\gamma = K + \gamma I_n$ and $C_\gamma := C + \gamma I$

Estimators via ERM

The estimators have the form $\widehat{G} = \widehat{S}^* W \widehat{Z}, \quad W \in \mathbb{R}^{n\times n}$

$$\min_{\widehat{G}\in \mathrm{HS}(\mathcal{H})} \widehat{\mathcal{R}}(\widehat{G}) + \gamma \|\widehat{G}\|^2_{\mathrm{HS}}$$

- Kernel Ridge Regression (KRR) $G_\gamma := C_\gamma^{-1} T$:
  $W = K_\gamma^{-1}$, with $K = (k(x_i, x_j))_{i,j=1}^n$, $K_\gamma = K + \gamma I_n$ and $C_\gamma := C + \gamma I$
- Principal Component Regression (PCR, aka Kernel-DMD) $G^{\mathrm{PCR}}_{r,\gamma} := [\![C_\gamma]\!]_r^\dagger T$:
  $W = [\![K_\gamma]\!]_r^\dagger$, where $[\![\cdot]\!]_r$ denotes r-truncated SVD

Estimators via ERM

The estimators have the form $\widehat{G} = \widehat{S}^* W \widehat{Z}, \quad W \in \mathbb{R}^{n \times n}$

$$\min_{\substack{\widehat{G} \in \mathrm{HS}(\mathcal{H}) \\ \mathrm{rank}(\widehat{G}) \leq r}} \widehat{\mathcal{R}}(\widehat{G}) + \gamma \|\widehat{G}\|_{\mathrm{HS}}^2$$

- Kernel Ridge Regression (KRR) $G_\gamma := C_\gamma^{-1} T$:
  $W = K_\gamma^{-1}$, with $K = (k(x_i, x_j))_{i,j=1}^n$, $K_\gamma = K + \gamma I_n$ and $C_\gamma := C + \gamma I$

- Principal Component Regression (PCR, aka Kernel-DMD) $G_{r,\gamma}^{\mathrm{PCR}} := [\![C_\gamma]\!]_r^\dagger T$:
  $W = [\![K_\gamma]\!]_r^\dagger$, where $[\![\cdot]\!]_r$ denotes r-truncated SVD

- Reduced Rank Regression (RRR) $G_{r,\gamma}^{\mathrm{RRR}} := C_\gamma^{-1/2} [\![C_\gamma^{-1/2} T]\!]_r$:
  $W = \sum_{i=1}^r u_i \otimes (K u_i)$ where u_i are the r leading eigenvectors of $L K u_i = \sigma_i^2 K_\gamma u_i$, normalized as $u_i^\top K K_\gamma u_i = 1$, and $L = (k(y_i, y_j))_{i,j=1}^n$

Estimators via ERM

The estimators have the form $\widehat{G} = \widehat{S}^* W \widehat{Z}, \quad W \in \mathbb{R}^{n\times n}$

$$\min_{\substack{\widehat{G}\in \mathrm{HS}(\mathcal{H}) \\ \mathrm{rank}(\widehat{G})\le r}} \widehat{\mathcal{R}}(\widehat{G}) + \gamma \|\widehat{G}\|^2_{\mathrm{HS}}$$

- Kernel Ridge Regression (KRR) $G_\gamma := C_\gamma^{-1} T$:
  $W = K_\gamma^{-1}$, with $K = (k(x_i, x_j))_{i,j=1}^n$, $K_\gamma = K + \gamma I_n$ and $C_\gamma := C + \gamma I$
- Principal Component Regression (PCR, aka Kernel-DMD) $G^{\mathrm{PCR}}_{r,\gamma} := [\![C_\gamma]\!]_r^\dagger T$:
  $W = [\![K_\gamma]\!]_r^\dagger$, where $[\![\cdot]\!]_r$ denotes r-truncated SVD
- Reduced Rank Regression (RRR) $G^{\mathrm{RRR}}_{r,\gamma} := C_\gamma^{-1/2} [\![C_\gamma^{-1/2} T]\!]_r$:
  $W = \sum_{i=1}^r u_i \otimes (K u_i)$ where u_i are the r leading eigenvectors of $LKu_i = \sigma_i^2 K_\gamma u_i$, normalized as $u_i^\top K K_\gamma u_i = 1$, and $L = (k(y_i, y_j))_{i,j=1}^n$

Theorem: Let $W = \sum_{i=1}^r u_i \otimes v_i$, then the modal decomposition of $\widehat{G}$ can be computed by solving an eigenvalue problem $(v_i^\top M u_j)_{i,j=1}^r \in \mathbb{R}^{r\times r}$, where $M = (k(x_i, y_j))_{i,j=1}^n$.

Learning KMD

Let $G \in \mathrm{HS}(\mathcal{H})$ be rank r and non-defective, then

$$G = \sum_{i=1}^{r} \lambda_i \, \psi_i \otimes \bar{\xi}_i, \quad G\psi_i = \lambda_i \psi_i, \quad G^* \xi_i = \overline{\lambda}_i \xi_i, \quad \langle \psi_i, \bar{\xi}_j \rangle_{\mathcal{H}} = \delta_{ij}, \; i, j \in [r],$$

and the mode decomposition of G is: $(G^t h)(x) = \sum_{i=1}^{r} \lambda_i^t \langle h, \bar{\xi}_i \rangle_{\mathcal{H}} \psi_i(x)$, $h \in \mathcal{H}$, $t \in \mathbb{N}$

Learning KMD

Let $G \in \mathrm{HS}(\mathcal{H})$ be rank r and non-defective, then

$$G = \sum_{i=1}^{r} \lambda_i \, \psi_i \otimes \bar{\xi}_i, \quad G\psi_i = \lambda_i \psi_i, \quad G^* \xi_i = \overline{\lambda}_i \xi_i, \quad \langle \psi_i, \bar{\xi}_j \rangle_{\mathcal{H}} = \delta_{ij}, \; i,j \in [r],$$

and the mode decomposition of G is: $(G^t h)(x) = \sum_{i=1}^{r} \lambda_i^t \langle h, \bar{\xi}_i \rangle_{\mathcal{H}} \psi_i(x)$, $h \in \mathcal{H}$, $t \in \mathbb{N}$

Theorem

(i) Forecasting can get increasingly harder for larger t:

$$\|\mathbb{E}[h(X_t) | X_0 = \cdot\,] - S_\pi G^t h\|_{L^2_\pi} \leq \underbrace{\|A_\pi S_\pi - S_\pi G\|}_{\text{operator norm error}} \Big(\textstyle\sum_{k=0}^{t-1} \|G^k\| \Big) \|h\|$$

Learning KMD

Let $G \in \mathrm{HS}(\mathcal{H})$ be rank r and non-defective, then

$$G = \sum_{i=1}^{r} \lambda_i \, \psi_i \otimes \bar{\xi}_i, \quad G\psi_i = \lambda_i \psi_i, \quad G^* \xi_i = \overline{\lambda}_i \xi_i, \quad \langle \psi_i, \bar{\xi}_j \rangle_{\mathcal{H}} = \delta_{ij}, \; i,j \in [r],$$

and the mode decomposition of G is: $(G^t h)(x) = \sum_{i=1}^{r} \lambda_i^t \langle h, \bar{\xi}_i \rangle_{\mathcal{H}} \psi_i(x)$, $h \in \mathcal{H}$, $t \in \mathbb{N}$

Theorem

(i) Forecasting can get increasingly harder for larger t:

$$\|\mathbb{E}[h(X_t) | X_0 = \cdot\,] - S_\pi G^t h\|_{L^2_\pi} \leq \underbrace{\|A_\pi S_\pi - S_\pi G\|}_{\text{operator norm error}} \Big(\textstyle\sum_{k=0}^{t-1} \|G^k\| \Big) \|h\|$$

(ii) The pseudo eigen-pair $(\lambda_i, S_\pi \psi_i)$ error may be looser than the operator norm error:

$$\|(A_\pi - \lambda_i I)^{-1}\|^{-1} \leq \frac{\|(A_\pi - \lambda_i I) S_\pi \psi_i\|}{\|S_\pi \psi_i\|} \leq \underbrace{\|A_\pi S_\pi - S_\pi G\|}_{\mathcal{E}(G)} \underbrace{\frac{\|\psi_i\|}{\|S_\pi \psi_i\|}}_{\eta(\psi_i)}$$

Learning KMD

Let $G \in \mathrm{HS}(\mathcal{H})$ be rank r and non-defective, then

$$G = \sum_{i=1}^{r} \lambda_i \, \psi_i \otimes \bar{\xi}_i, \qquad G\psi_i = \lambda_i \psi_i, \qquad G^* \xi_i = \bar{\lambda}_i \xi_i, \qquad \langle \psi_i, \bar{\xi}_j \rangle_{\mathcal{H}} = \delta_{ij}, \; i,j \in [r],$$

and the mode decomposition of G is: $(G^t h)(x) = \sum_{i=1}^{r} \lambda_i^t \langle h, \bar{\xi}_i \rangle_{\mathcal{H}} \psi_i(x)$, $h \in \mathcal{H}$, $t \in \mathbb{N}$

Theorem

(i) Forecasting can get increasingly harder for larger t:

$$\|\mathbb{E}[h(X_t) | X_0 = \cdot\,] - S_\pi G^t h\|_{L^2_\pi} \leq \underbrace{\|A_\pi S_\pi - S_\pi G\|}_{\text{operator norm error}} \Big(\textstyle\sum_{k=0}^{t-1} \|G^k\| \Big) \|h\|$$

(ii) The pseudo eigen-pair $(\lambda_i, S_\pi \psi_i)$ error may be looser than the operator norm error:

$$\|(A_\pi - \lambda_i I)^{-1}\|^{-1} \leq \frac{\|(A_\pi - \lambda_i I) S_\pi \psi_i\|}{\|S_\pi \psi_i\|} \leq \underbrace{\|A_\pi S_\pi - S_\pi G\|}_{\mathcal{E}(G)} \underbrace{\frac{\|\psi_i\|}{\|S_\pi \psi_i\|}}_{\eta(\psi_i)}$$

To get grantees for KMD one needs to control operator norm error and metric distortion!

Key players: operator norm error and metric distortion

- **Metric distortion:** Let $\widehat{G} \in \mathrm{HS}_r(\mathcal{H})$. Then for all $i \in [r]$

$$\frac{1}{\sqrt{\|C\|}} \leq \eta(\widehat{\psi}_i) \leq \frac{|\widehat{\lambda}_i| \, \mathrm{cond}(\widehat{\lambda}_i) \wedge \|\widehat{G}\|}{\sigma^+_{\min}(S_\pi \widehat{G})},$$

where $\mathrm{cond}(\widehat{\lambda}_i) := \|\widehat{\xi}_i\| \|\widehat{\psi}_i\| / |\langle \widehat{\psi}_i, \widehat{\xi}_i \rangle_{\mathcal{H}}|$ is the condition number of $\widehat{\lambda}_i$

Key players: operator norm error and metric distortion

- **Metric distortion:** Let $\widehat{G} \in \mathrm{HS}_r(\mathcal{H})$. Then for all $i \in [r]$

$$\frac{1}{\sqrt{\|C\|}} \leq \eta(\widehat{\psi}_i) \leq \frac{|\widehat{\lambda}_i| \, \mathrm{cond}(\widehat{\lambda}_i) \wedge \|\widehat{G}\|}{\sigma_{\min}^{+}(S_\pi \widehat{G})},$$

where $\mathrm{cond}(\widehat{\lambda}_i) := \|\widehat{\xi}_i\| \|\widehat{\psi}_i\| / |\langle \widehat{\psi}_i, \widehat{\xi}_i \rangle_{\mathcal{H}}|$ is the condition number of $\widehat{\lambda}_i$

- **Operator norm error:** to analyze it we use the following decomposition

$$\mathcal{E}(\widehat{G}) \leq \underbrace{\|[I - P_{\mathcal{H}}] A_\pi S_\pi\|}_{\text{kernel selection bias}} + \underbrace{\|P_{\mathcal{H}} A_\pi S_\pi - S_\pi G_\gamma\|}_{\text{regularization bias}} + \underbrace{\|S_\pi (G_\gamma - G)\|}_{\text{rank reduction bias}} + \underbrace{\|S_\pi (G - \widehat{G})\|}_{\text{estimator's variance}},$$

where $G_\gamma := C_\gamma^{-1} T = \arg\min_{G \in \mathrm{HS}(\mathcal{H})} \mathcal{R}(G) + \gamma \|G\|_{\mathrm{HS}}^2$, and G being is the population version of the empirical estimator $\widehat{G}$.

Assumptions for deriving the learning bounds

(BC) Boundedness of the kernel. There exists $c_{\mathcal{H}} > 0$ such that $\operatorname*{ess\,sup}_{x \sim \pi} \|\phi(x)\|^2 \leq c_{\mathcal{H}}$

Assumptions for deriving the learning bounds

(BC) Boundedness of the kernel. There exists $c_{\mathcal{H}} > 0$ such that $\operatorname*{ess\,sup}_{x \sim \pi} \|\phi(x)\|^2 \leq c_{\mathcal{H}}$

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

Assumptions for deriving the learning bounds

(BC) Boundedness of the kernel. There exists $c_{\mathcal{H}} > 0$ such that $\operatorname*{ess\,sup}_{x\sim\pi} \|\phi(x)\|^2 \leq c_{\mathcal{H}}$

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

(RC) Regularity of A_π. For some $\alpha \in (0,2]$ there exists $a > 0$ such that $TT^* \preceq a^2 C^{1+\alpha}$.

Assumptions for deriving the learning bounds

(BC) Boundedness of the kernel. There exists $c_{\mathcal{H}} > 0$ such that $\operatorname*{ess\,sup}_{x\sim\pi} \|\phi(x)\|^2 \leq c_{\mathcal{H}}$

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

(RC) Regularity of A_π. For some $\alpha \in (0, 2]$ there exists $a > 0$ such that $TT^* \preceq a^2 C^{1+\alpha}$.

- (RC) is weaker than the existing source condition (SRC) used for CME analysis that relies on the interpolation spaces, i.e. $\mathrm{Im}(A_\pi S_\pi) \subseteq \mathrm{Im}(S_\pi\, C^{(\alpha-1)/2})$

Assumptions for deriving the learning bounds

(BC) Boundedness of the kernel. There exists $c_{\mathcal{H}} > 0$ such that $\operatorname*{ess\,sup}_{x\sim\pi} \|\phi(x)\|^2 \leq c_{\mathcal{H}}$

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

(RC) Regularity of A_π. For some $\alpha \in (0, 2]$ there exists $a > 0$ such that $TT^* \preceq a^2 C^{1+\alpha}$.

- (RC) is weaker than the existing source condition (SRC) used for CME analysis that relies on the interpolation spaces, i.e. $\mathrm{Im}(A_\pi S_\pi) \subseteq \mathrm{Im}(S_\pi\, C^{(\alpha-1)/2})$

- For example, with Gaussian RKHS ($\beta \to 0$), (SRC) does not hold for any $\alpha \in (0, 2]$, while if $A_\pi^* = A_\pi$ assumption (RC) holds true for at least $\alpha = 1$.

Error Learning Bounds

Theorem (Operator norm error)

Let A_π be an operator such that $\sigma_r(A_\pi S_\pi) > \sigma_{r+1}(A_\pi S_\pi) \geq 0$ for some $r \in \mathbb{N}$. Let (SD) and (RC) hold for some $\beta \in (0,1]$ and $\alpha \in [1,2]$, respectively, and let $\mathrm{cl}(\mathrm{Im}(S_\pi)) = L^2_\pi(\mathcal{X})$. Given $\delta \in (0,1)$ let

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

Then, there exists a constant $c > 0$, such that for large enough $n \geq r$ and every $i \in [r]$, with probability at least $1-\delta$ in the i.i.d. draw of $(x_i, y_i)_{i=1}^n$ from ρ

$$\mathcal{E}(\widehat{G}_{\mathrm{RRR}}) \leq \sigma_{r+1}(A_\pi S_\pi) + c\,\varepsilon^\star_n \ln \delta^{-1}$$

and, assuming that $\sigma_r(S_\pi) > \sigma_{r+1}(S_\pi)$,

$$\mathcal{E}(\widehat{G}_{\mathrm{PCR}}) \leq \sigma_{r+1}(S_\pi) + c\,\varepsilon^\star_n \ln \delta^{-1}.$$

Error Learning Bounds

Theorem (Operator norm error)

Let A_π *be an operator such that* $\sigma_r(A_\pi S_\pi) > \sigma_{r+1}(A_\pi S_\pi) \geq 0$ *for some* $r \in \mathbb{N}$. *Let (SD) and (RC) hold for some* $\beta \in (0,1]$ *and* $\alpha \in [1,2]$, *respectively, and let* $\mathrm{cl}(\mathrm{Im}(S_\pi)) = L^2_\pi(\mathcal{X})$. *Given* $\delta \in (0,1)$ *let*

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

Then, there exists a constant $c > 0$, *such that for large enough* $n \geq r$ *and every* $i \in [r]$, *with probability at least* $1-\delta$ *in the i.i.d. draw of* $(x_i, y_i)_{i=1}^n$ *from* ρ

$$\mathcal{E}(\widehat{G}_{\mathrm{RRR}}) \leq \sigma_{r+1}(A_\pi S_\pi) + c\,\varepsilon_n^\star \ln \delta^{-1}$$

and, assuming that $\sigma_r(S_\pi) > \sigma_{r+1}(S_\pi)$,

$$\mathcal{E}(\widehat{G}_{\mathrm{PCR}}) \leq \sigma_{r+1}(S_\pi) + c\,\varepsilon_n^\star \ln \delta^{-1}.$$

Moreover, the rate matches the minimax lower bound for the operator norm error when learning finite rank A_π, $r \geq 2$,

$$\mathcal{E}(\widehat{G}) \geq c\,\delta^q\,\varepsilon_n^\star.$$

Koopman spectra for time-reversal invariant processes

Example (Langevin Dynamics)

Let $\mathcal{X} = \mathbb{R}^d$ and let $\beta > 0$. The (overdamped) Langevin equation driven by a potential $U : \mathbb{R}^d \to \mathbb{R}$ is given by

$$dX_t = -\nabla U(X_t)dt + \sqrt{2\beta^{-1}}dW_t,$$

where W_t is a Wiener process. The invariant measure of this process is the *Boltzman distribution* $\pi(dx) \propto e^{-\beta U(x)}dx$, and the associated Koopman operator is self-adjoint.

Koopman spectra for time-reversal invariant processes

Example (Langevin Dynamics)

Let $\mathcal{X} = \mathbb{R}^d$ and let $\beta > 0$. The (overdamped) Langevin equation driven by a potential $U : \mathbb{R}^d \to \mathbb{R}$ is given by

$$dX_t = -\nabla U(X_t)dt + \sqrt{2\beta^{-1}}dW_t,$$

where W_t is a Wiener process. The invariant measure of this process is the *Boltzman distribution* $\pi(dx) \propto e^{-\beta U(x)}dx$, and the associated Koopman operator is self-adjoint.

- Koopman operator for time-reversal invariant processes is self-adjoint, i.e. $A_\pi^* = A_\pi$.

Koopman spectra for time-reversal invariant processes

Example (Langevin Dynamics)

Let $\mathcal{X} = \mathbb{R}^d$ and let $\beta > 0$. The (overdamped) Langevin equation driven by a potential $U : \mathbb{R}^d \to \mathbb{R}$ is given by

$$dX_t = -\nabla U(X_t)dt + \sqrt{2\beta^{-1}}dW_t,$$

where W_t is a Wiener process. The invariant measure of this process is the *Boltzman distribution* $\pi(dx) \propto e^{-\beta U(x)}dx$, and the associated Koopman operator is self-adjoint.

- Koopman operator for time-reversal invariant processes is self-adjoint, i.e. $A_\pi^* = A_\pi$.
- If additionally we assume compactness of A_π (e.g. if $p(x, \cdot) \ll \pi$, for all $x \in \mathcal{X}$), then

$$A_\pi = \textstyle\sum_{i\in\mathbb{N}} \mu_i \, f_i \otimes f_i,$$

  where $(\mu_i, f_i)_{i\in\mathbb{N}} \subseteq \mathbb{R} \times L^2_\pi(\mathcal{X})$ are Koopman eigenpairs, i.e. $A_\pi f_i = \mu_i f_i$. Moreover, $\lim_{i\to\infty} \mu_i = 0$ and $\{f_i\}_{i\in\mathbb{N}}$ form a complete orthonormal system of $L^2_\pi(\mathcal{X})$.

Estimation of Koopman spectra in self-adjoint case

- Let $(\widehat{\lambda}_i, \widehat{\psi}_i)_{i=1}^r$ be its eigen-pairs a rank r estimator $\widehat{G} \in \mathrm{HS}\,(\mathcal{H})$ of A_π, i.e. $\widehat{G}\widehat{\psi}_i = \widehat{\lambda}_i\,\widehat{\psi}_i$.

Estimation of Koopman spectra in self-adjoint case

- Let $(\widehat{\lambda}_i, \widehat{\psi}_i)_{i=1}^r$ be its eigen-pairs a rank r estimator $\widehat{G} \in \mathrm{HS}(\mathcal{H})$ of A_π, i.e. $\widehat{G}\widehat{\psi}_i = \widehat{\lambda}_i \widehat{\psi}_i$.
- To compare $\widehat{\psi}_i$ with the corresponding true Koopman eigenfunction f_i, using S_π , we inject $\widehat{\psi}_i$ in $L^2_\pi(\mathcal{X})$ to define the normalized estimated eigenfunction

$$\widehat{f}_i = S_\pi \widehat{\psi}_i \,/\, \|S_\pi \widehat{\psi}_i\|, \; i \in [r].$$

Estimation of Koopman spectra in self-adjoint case

- Let $(\widehat{\lambda}_i, \widehat{\psi}_i)_{i=1}^r$ be its eigen-pairs a rank r estimator $\widehat{G} \in \mathrm{HS}(\mathcal{H})$ of A_π, i.e. $\widehat{G}\widehat{\psi}_i = \widehat{\lambda}_i \widehat{\psi}_i$.
- To compare $\widehat{\psi}_i$ with the corresponding true Koopman eigenfunction f_i, using S_π , we inject $\widehat{\psi}_i$ in $L^2_\pi(\mathcal{X})$ to define the normalized estimated eigenfunction

$$\widehat{f}_i = S_\pi \widehat{\psi}_i \,/\, \|S_\pi \widehat{\psi}_i\|, \; i \in [r].$$

- Using the classical Davis-Kahan spectral perturbation result we get

$$|\widehat{\lambda}_i - \mu_i| \leq \|(\widehat{\lambda}_i I - A_\pi)^{-1}\|^{-1} \leq \mathcal{E}(\widehat{G})\, \eta(\widehat{\psi}_i), \quad \text{and}$$

$$\|\widehat{f}_i - f_i\|^2 \leq \frac{2|\widehat{\lambda}_i - \mu_i|}{[\mathrm{gap}_i(A_\pi) - |\widehat{\lambda}_i - \mu_i|]_+},$$

where $\mathrm{gap}_i(A_\pi) = \min_{j \neq j} |\mu_j - \mu_j|$.

Estimation of Koopman spectra in self-adjoint case

- Let $(\widehat{\lambda}_i, \widehat{\psi}_i)_{i=1}^r$ be its eigen-pairs a rank r estimator $\widehat{G} \in \mathrm{HS}\,(\mathcal{H})$ of A_π, i.e. $\widehat{G}\widehat{\psi}_i = \widehat{\lambda}_i\,\widehat{\psi}_i$.
- To compare $\widehat{\psi}_i$ with the corresponding true Koopman eigenfunction f_i, using S_π , we inject $\widehat{\psi}_i$ in $L^2_\pi(\mathcal{X})$ to define the normalized estimated eigenfunction

$$\widehat{f}_i = S_\pi \widehat{\psi}_i \,/\, \|S_\pi \widehat{\psi}_i\|,\; i \in [r].$$

- Using the classical Davis-Kahan spectral perturbation result we get

$$|\widehat{\lambda}_i - \mu_i| \leq \|(\widehat{\lambda}_i I - A_\pi)^{-1}\|^{-1} \leq \mathcal{E}(\widehat{G})\,\eta(\widehat{\psi}_i), \quad \text{and}$$

$$\|\widehat{f}_i - f_i\|^2 \leq \frac{2|\widehat{\lambda}_i - \mu_i|}{[\mathrm{gap}_i(A_\pi) - |\widehat{\lambda}_i - \mu_i|]_+},$$

where $\mathrm{gap}_i(A_\pi) = \min_{j\neq j}|\mu_j - \mu_j|$.

- **Spuriousness** of spectra can arise purely from the learning problem, i.e.

"well learned" operator (small error) but "badly learned" spectra (eigenvalues far apart)

Spectral Learning Bounds

Theorem (Spectral bounds for self-adjoint Koopman)

Let A_π be a compact self-adjoint operator. Under the assumptions of the previous Theorem, there exists a constant $c > 0$, depending only on $\mathcal{H}$, such that for every $\delta \in (0,1)$, for every large enough $n \geq r$ and every $i \in [r]$ with probability at least $1-\delta$ in the i.i.d. draw of $(x_i, y_i)_{i=1}^n$ from ρ

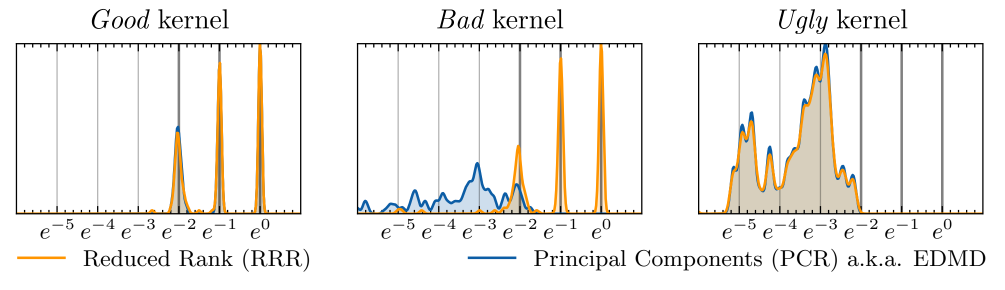
$$|\widehat{\lambda}_i - \mu_{j(i)}| \leq \begin{cases} \frac{2\sigma_{r+1}(A_\pi S_\pi)}{\sigma_r(A_\pi S_\pi)} + c\,\varepsilon_n^\star \ln \delta^{-1} & \text{if } \widehat{G} = \widehat{G}^{\mathrm{RRR}}_{r,\gamma}, \\ \frac{2\sigma_{r+1}(S_\pi)}{[\sigma_r(A_\pi S_\pi) - \sigma^\alpha_{r+1}(S_\pi)]_+} + c\,\varepsilon_n^\star \ln \delta^{-1} & \text{if } \widehat{G} = \widehat{G}^{\mathrm{PCR}}_{r,\gamma}. \end{cases}$$

Moreover, $|\widehat{\lambda}_i - \mu_{j(i)}| \leq s_i(\widehat{G}) + +c\,\varepsilon_n^\star \ln \delta^{-1}$, where the empirical bias is given by

$$s_i(\widehat{G}) := \begin{cases} \widehat{\eta}_i\, \sigma_{r+1}(\widehat{C}^{-1/2}\widehat{T}), & \widehat{G} = \widehat{G}^{\mathrm{RRR}}_{r,\gamma}, \\ \widehat{\eta}_i \sqrt{\sigma_{r+1}(\widehat{C})}, & \widehat{G} = \widehat{G}^{\mathrm{PCR}}_{r,\gamma}. \end{cases}$$

Experiments

Example: Choice of the kernel



PCR vs. RRR in estimating slow dynamics of 1D Ornstein–Uhlenbeck process

$$X_t = e^{-1} X_{t-1} + \sqrt{1 - e^{-2}}\, \epsilon_t,$$

where $\{\epsilon_t\}_{t \geq 1}$ are independent standard Gaussians.

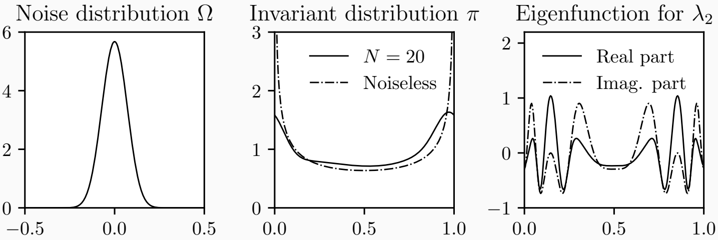
We use three different kernels over 50 independent trials. Vertical lines correspond to Koopman eigenvalues. The *good* kernel is such that its $\mathcal{H}$ corresponds to the leading eigenspace of the Koopman operator, while the other two use permuted eigenfunctions to distort the metric and introduce slow (*bad* kernel) and fast (*ugly* kernel) spectral decay of the covariance.

Example: Noisy Logistic Map

Let $F(x) := 4x(1-x)$ over $\mathcal{X} = [0,1]$ and consider the discrete dynamical system

$$x_{t+1} = (F(x_t) + \xi_t) \mod 1,$$

where ξ_t are i.i.d. with law $\Omega(d\xi) \propto \cos^N(\pi\xi)d\xi$, N *even*



For this system we are able to evaluate the spectral decomposition of A_π: $\mathrm{rank}(A_\pi) = N+1$ and the eigenvalues decay fast: $\lambda_1 = 1$, $\lambda_{2,3} = -0.193 \pm 0.191i$, and $|\lambda_{4,5}| \approx 0.027$.

Example: Noisy Logistic Map

Experimental setting: 10^4 training points, 500 test points, 100 repetitions

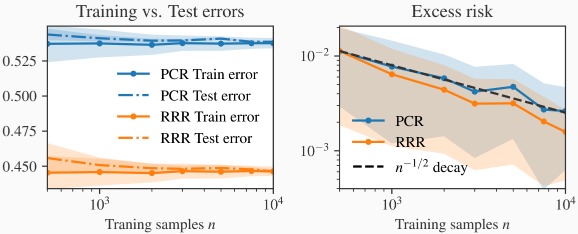
Estimator	Training error	Test error
PCR	0.2 ± 0.003	0.18 ± 0.00051
RRR	0.13 ± 0.002	$\mathbf{0.13 \pm 0.00032}$
KRR	$\mathbf{0.032 \pm 0.00057}$	$\mathbf{0.13 \pm 0.00068}$

Example: Noisy Logistic Map

Experimental setting: 10^4 training points, 500 test points, 100 repetitions

Estimator	Training error	Test error
PCR	0.2 ± 0.003	0.18 ± 0.00051
RRR	0.13 ± 0.002	$\mathbf{0.13 \pm 0.00032}$
KRR	$\mathbf{0.032 \pm 0.00057}$	$\mathbf{0.13 \pm 0.00068}$

- Empirically we verify bounds!



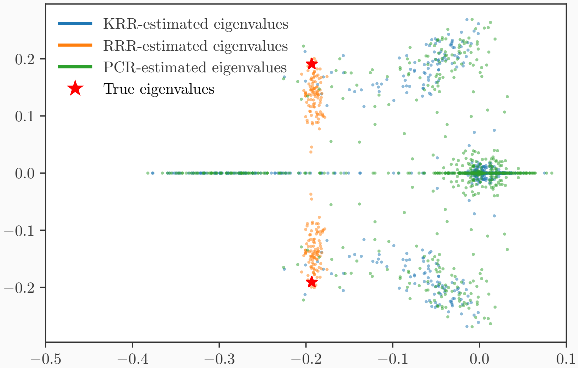
Example: Noisy Logistic Map

Experimental setting: 10^4 training points, 500 test points, 100 repetitions

Estimator	Training error	Test error	$\lvert\lambda_1 - \hat{\lambda}_1\rvert / \lvert\lambda_1\rvert$	$\lvert\lambda_{2,3} - \hat{\lambda}_{2,3}\rvert / \lvert\lambda_{2,3}\rvert$
PCR	0.2 ± 0.003	0.18 ± 0.00051	$9.6 \cdot 10^{-5} \pm 7.2 \cdot 10^{-5}$	0.85 ± 0.03
RRR	0.13 ± 0.002	$\mathbf{0.13 \pm 0.00032}$	$5.1 \cdot 10^{-6} \pm 3.8 \cdot 10^{-6}$	$\mathbf{0.16 \pm 0.1}$
KRR	$\mathbf{0.032 \pm 0.00057}$	$\mathbf{0.13 \pm 0.00068}$	$\mathbf{7.9 \cdot 10^{-7} \pm 5.7 \cdot 10^{-7}}$	0.48 ± 0.17

- Empirically we verify bounds!
- $\lambda_1 = 1$ (corresponding to the *equilibrium mode*) is well approximated by all estimators
- RRR always outperforms PCR and it best estimates the non-trivial eigenvalues $\lambda_{2,3}$

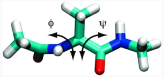
Estimated eigenvalues over 100 different independent datasets

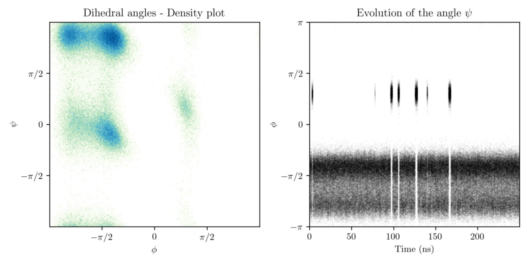


Example: Koopman Operator & Molecular Dynamics

Simulation of the molecule Alanine dipeptide from the Computational Molecular Biology Group, Freie Universität Berlin:

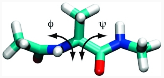
- dynamics governed by the Langevin equation is Markovian
- exists an invariant measure called Boltzmann distribution
- equations are time-reversal-invariant, so $A_\pi = A_\pi^*$



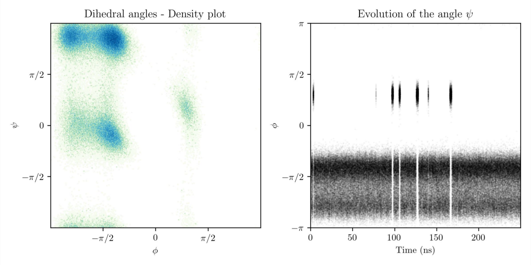


Example: Koopman Operator & Molecular Dynamics

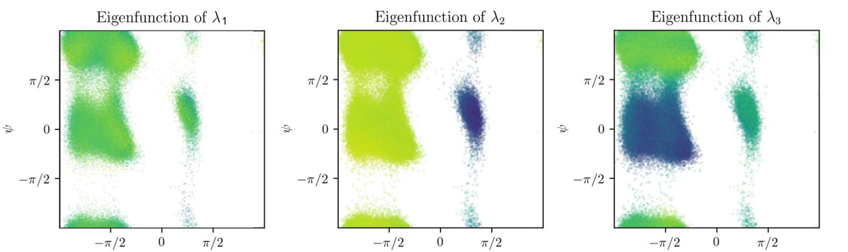
Simulation of the molecule Alanine dipeptide from the Computational Molecular Biology Group, Freie Universität Berlin:



- dynamics governed by the Langevin equation is Markovian
- exists an invariant measure called Boltzmann distribution
- equations are time-reversal-invariant, so $A_\pi = A_\pi^*$



The estimated evals $\lambda_1 = 0.9992$, $\lambda_2 = 0.9177$, $\lambda_3 = 0.4731$, $\lambda_4 = -0.0042$ and $\lambda_5 = -0.0252$.



Example: Koopman Operator & Molecular Dynamics

- In this example we show that minimizing the empirical spectral bias over a validation dataset, is also a good criterion for Koopman model selection.

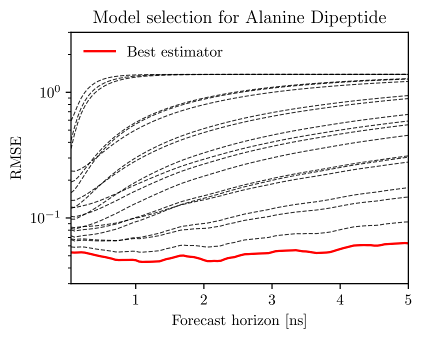
Example: Koopman Operator & Molecular Dynamics

- In this example we show that minimizing the empirical spectral bias over a validation dataset, is also a good criterion for Koopman model selection.
- We trained 19 RRR estimators each corresponding to a different kernel and then we evaluated the forecasting RMSE over 5000 validation points from 2000 initial conditions drawn from a test dataset.

Example: Koopman Operator & Molecular Dynamics

- In this example we show that minimizing the empirical spectral bias over a validation dataset, is also a good criterion for Koopman model selection.
- We trained 19 RRR estimators each corresponding to a different kernel and then we evaluated the forecasting RMSE over 5000 validation points from 2000 initial conditions drawn from a test dataset.

- Forecasting RMSE shows how the best model according to the empirical spectral bias metric also attains the best forecasting performances by a large margin.



Example: Koopman Operator with "Deep" Kernels

- In computer vision, kernels defined from neural-network feature maps outperform classical ones

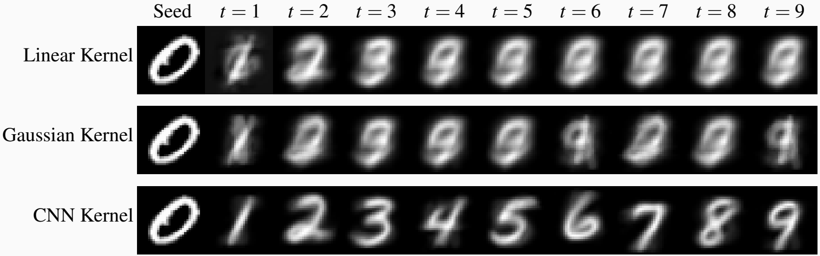
Example: Koopman Operator with "Deep" Kernels

- In computer vision, kernels defined from neural-network feature maps outperform classical ones

- We compare Linear, Gaussian and *Convolutional Neural Network (CNN)* kernels, the latter being

$$k_{\boldsymbol{w}}(x, x') := \langle \phi_{\boldsymbol{w}}(x), \phi_{\boldsymbol{w}}(x') \rangle$$

  where $\phi_{\boldsymbol{w}}$ is the last layer of a pretrained CNN classifier. Training data size $= 1000$



Conv2d(1,16; 5) → ReLU → MaxPool(2) → Conv2d(16,32; 5) → ReLU → MaxPool(2) → Dense(1568, 10)

Deep Learning of a good RKHS

Deep Projection Networks

- **What is a good RKHS?**
  dominant Koopman efuns captured, no kernel selection bias and no metric distorsion

$$P_{\mathcal{H}} A_\pi P_{\mathcal{H}} \approx A_\pi, \quad \|[I - P_{\mathcal{H}}] A_\pi S_\pi\| \rightsquigarrow 0 \quad \text{and} \quad \eta(\psi) = \|\psi\| \,/\, \|C^{1/2}\psi\| \rightsquigarrow 1$$

Deep Projection Networks

- **What is a good RKHS?**
  dominant Koopman efuns captured, no kernel selection bias and no metric distorsion

$$P_{\mathcal{H}} A_\pi P_{\mathcal{H}} \approx A_\pi, \quad \|[I - P_{\mathcal{H}}]A_\pi S_\pi\| \rightsquigarrow 0 \quad \text{and} \quad \eta(\psi) = \|\psi\| \,/\, \|C^{1/2}\psi\| \rightsquigarrow 1$$

- The idea is to parameterize two feature vectors one for input and one for the output:

$$\phi_w(x) := [\phi_{w,1}(x), \ldots, \phi_{w,\ell}(x)] \in \mathbb{R}^\ell \text{ and } \phi_{w'}(y) := [\phi_{w',1}(y), \ldots, \phi_{w',\ell}(y)] \in \mathbb{R}^\ell$$

and then, using covariance operators

$$C_X^w = \mathbb{E}\phi_w(X) \otimes \phi_w(X),\; C_{XY}^{ww'} = \mathbb{E}\phi_w(X) \otimes \phi_{w'}(Y) \text{ and } C_Y^{w'} = \mathbb{E}\phi_{w'}(Y) \otimes \phi_{w'}(Y),$$

maximize the regularized score

$$\max_{w,w'} \quad \underbrace{\frac{\|C_{XY}^{ww'}\|_{\mathrm{HS}}^2}{\|C_X^w\|\|C_Y^{w'}\|}}_{\leq \|P_{\mathcal{H}_w} A_\pi P_{\mathcal{H}_{w'}}\|_{\mathrm{HS}}^2} - \gamma \underbrace{\left(\|C_X^w - I\|_{\mathrm{HS}}^2 + \|C_Y^{w'} - I\|_{\mathrm{HS}}^2\right)}_{\text{reducing the metric distortion}}$$

Challenges & open problems

Thank You!

Trajectory data

- With notion of beta mixing coefficients:

$$\beta_{\mathbf{X}}(\tau) = \sup_{B \in \Sigma \otimes \Sigma} \left| \mu_{\{1,1+\tau\}}(B) - \mu_{\{1\}} \times \mu_{\{1\}}(B) \right|$$

we prove that for $B \in \Sigma_{[1:m]}$ $\left| \mu_{[1:m]}(B) - \mu^m_{\{1\}}(B) \right| \leq (m-1)\,\beta_{\mathbf{X}}(1)$, and derive

- **Lemma 1** Let $\mathbf{X}$ be strictly stationary with values in a normed space $(\mathcal{X}, \|\cdot\|)$, and assume $n = 2m\tau$ for $\tau, m \in \mathbb{N}$. Moreover, let $Z_1, \ldots, Z_m$ be m independent copies of $Z_1 = \sum_{i=1}^{\tau} X_i$. Then for $s > 0$

$$\mathbb{P}\left\{ \left\| \sum_{i=1}^{n} X_i \right\| > s \right\} \leq 2\,\mathbb{P}\left\{ \left\| \sum_{j=1}^{m} Z_j \right\| > \frac{s}{2} \right\} + 2\,(m-1)\,\beta_{\mathbf{X}}(\tau).$$

- We generalize Prop. 2 as

Proposition 3: Let $\delta > (m-1)\beta_{\mathbf{X}}(\tau - 1)$. With probability at least $1 - \delta$ in the draw $x_1 \sim \pi, x_i \sim p(x_{i-1}, \cdot)$, $i \in [2{:}n]$,

$$\|\widehat{T} - T\| \leq \frac{48}{m} \ln \frac{4m\tau}{\delta - (m-1)\,\beta_{\mathbf{X}}(\tau - 1)} + 12\sqrt{\frac{2\,\|C\|}{m} \ln \frac{4m\tau}{\delta - (m-1)\,\beta_{\mathbf{X}}(\tau - 1)}}.$$