Learning Representations of Markov Processes

ISTITUTO ITALIANO
DI TECNOLOGIA

ÉCOLE POLYTECHNIQUE


Massimiliano Pontil


Karim Lounici


Pietro Novelli

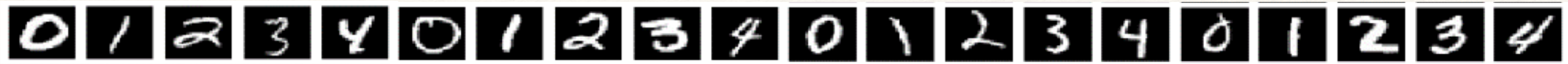
… and
Giacomo Turri,
Gregoire Pacreau,
Riccardo Grazzi,
Hélène Halconruy,
Timothée Devergne,
Michelle Parinello.

A short story on Machine Learning Theory meets Dynamical Systems Modelling

Funded by the European Union NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
FAIR Future Artificial Intelligence Research
RAISE

CDC2024 Workshop on
Data-driven modelling, analysis, and control using the Koopman operator

An example of learning to write digits



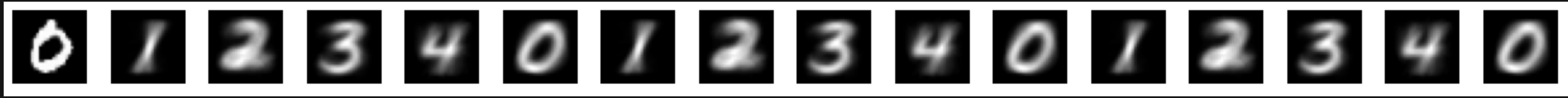
- Learner is given visual examples of digits 0, 1, 2, 3, 4, 0, 1, 2, …
- The task is, when presented another (unseen) example, to produce the images that continue the sequence
- Dataset: MNIST
- Learning task is unsupervised (we don’t see the labels)

- Problem can be formulated as a stochastic process (dynamical system)
- Ambient dimension is large (~784) but the effective dimension is small (cyclic order of 5 classes)
- Challenges:
  1) we don’t know the data distributions
  2) we don’t know the transition rule
  (*its’ a non deterministic one!*)

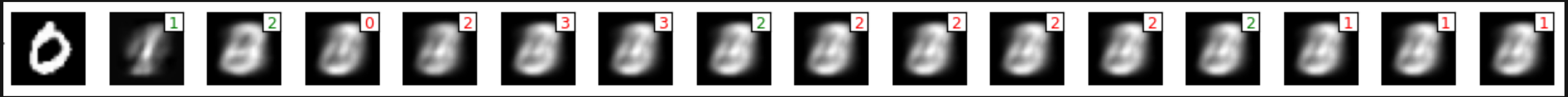
An example of learning to write digits

- Let's solve it with linear vector valued regression (aka Galerkin projection)

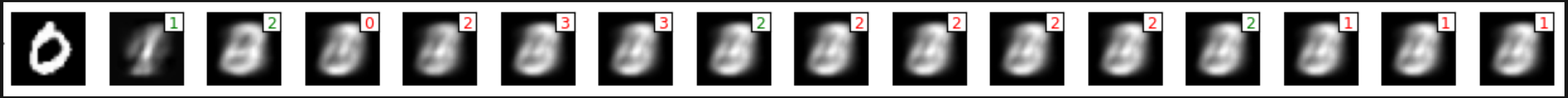
✓ Optimal solution, i.e. $\mathbb{E}[X_{t+1} \mid X_t]$



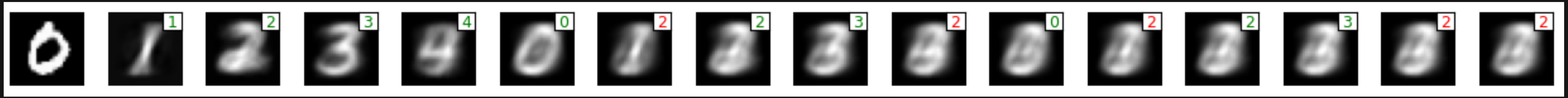
✓ Linear regression ~ linear dynamics?!



✓ RBF Kernel regression ~ linear dynamics in the RKHS?!



✓ CNN classifier features regression ~ linear in a representation space?!



Regression vs Operator Regression

- Regression: given $(X, Y) \sim \mu_{X,Y}$ learn $f\colon \mathcal{X} \to \mathcal{Y}$ s.t. $Y = f(X)$
  - Optimal solution w.r.t. MSE is the regression function $\mathbb{E}[Y \,|\, X = \cdot\,]$
  - So, we just learn the conditional mean. Can we learn distribution?

- Operator perspective: let $E_{Y|X}\colon \mathscr{L}^2_{\mu_Y}(\mathcal{Y}) \to \mathscr{L}^2_{\mu_X}(\mathcal{X})$ s.t. $E_{Y|X} f = \mathbb{E}[f(Y) \,|\, X = \cdot\,]$
  - Applying $E_{Y|X}$ to characteristic functions of sets, we obtain probabilities
  - Solving the <u>linear operator regression</u> problem we can predict <u>conditional probability distributions</u>!

Reminder on Transfer Operators

Stochastic Koopman

Consider time-homogenous Markov process $(X_t)_{t\in\mathbb{T}} \subseteq \mathcal{X}$, $\quad X_t \sim \mu_t$, i.e. $\mathbb{P}\left[X_{s+t}\,|\,X_{\leq s}\right] = \mathbb{P}\left[X_{s+t}\,|\,X_s\right]$ independent of $s, s+t \in \mathbb{T}$, which is described

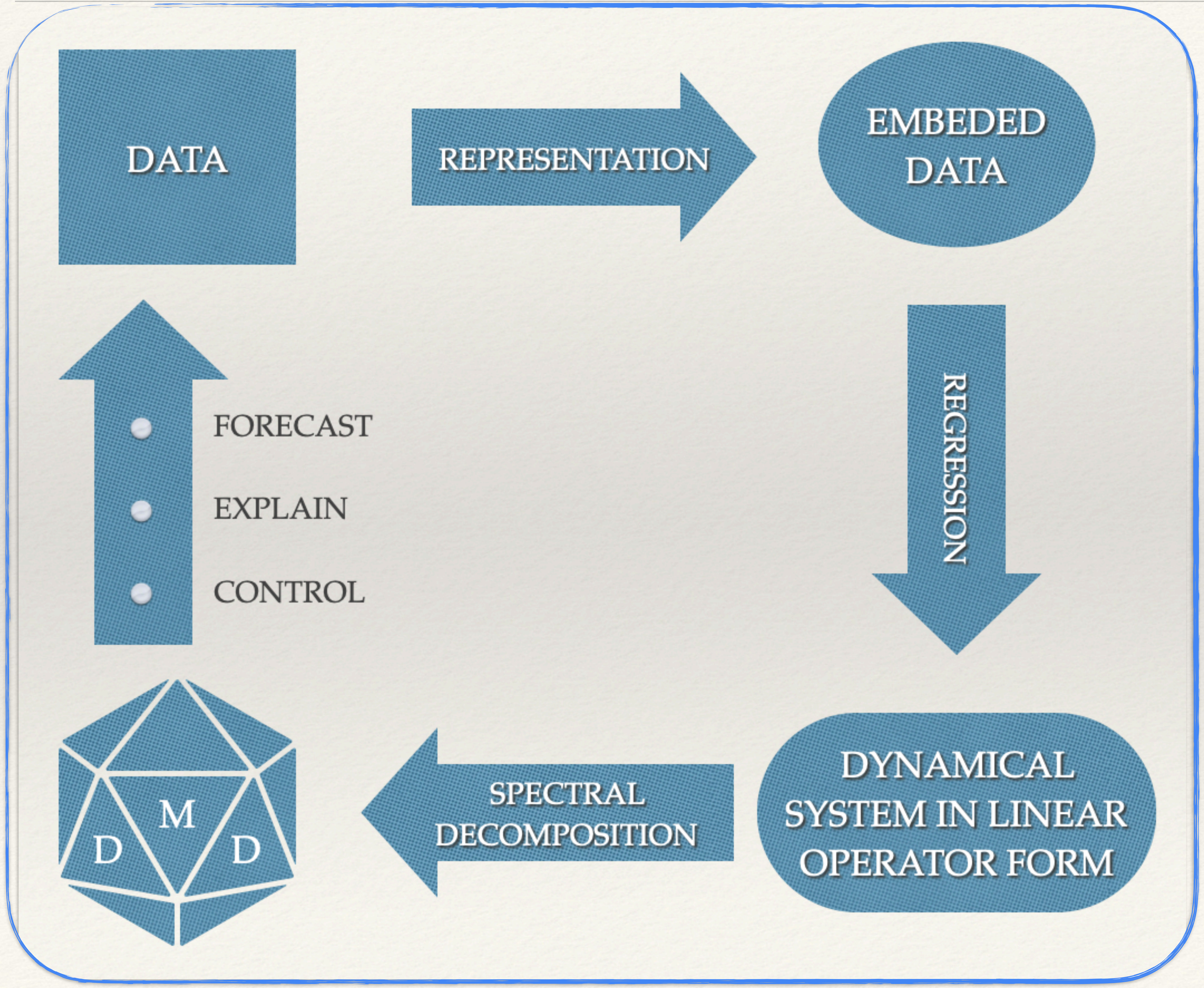
- in discrete time ($\mathbb{T}=\mathbb{N}$ & $s=1$) by transfer operators $E_{X_{s+1}\,|\,X_s} = \mathbb{E}[[\,\cdot\,](X_{s+1})\,|\,X_s]$
- in continous time ($\mathbb{T}=\mathbb{R}_+$) by TO semigroup $(E_{X_{s+t}\,|\,X_s})_{t\geq 0}$
- and when is stationary $(\forall t\in\mathbb{T})(\mu_t=\pi)$, by linear dynamical system in a function space, i.e. for $A_t = E_{X_{s+t}|X_s}\colon \mathcal{L}^2_\pi(\mathcal{X}) \to \mathcal{L}^2_\pi(\mathcal{X})$ and $q_t = d\mu_t/d\pi \in L^2_\pi(\mathcal{X})$

$$\boxed{q_t = (A_1^*)^t q_0,\ t\in\mathbb{N} \quad \text{and} \quad q_t = e^{tL^*} q_0,\ t\in\mathbb{R}_+, \quad \text{where} \quad L = \lim_{t\to 0^+}(A_t - I)/t}$$

since

$$\langle q_{s+t}, f\rangle_{L^2_\pi(\mathcal{X})} = \mathbb{E}[f(X_{s+t})] = \mathbb{E}[\mathbb{E}[f(X_{s+t})\,|\,X_s]] = \langle q_s, A_t f\rangle_{L^2_\pi(\mathcal{X})} = \langle A_t^* q_s, f\rangle_{L^2_\pi(\mathcal{X})}$$

General learning pipeline



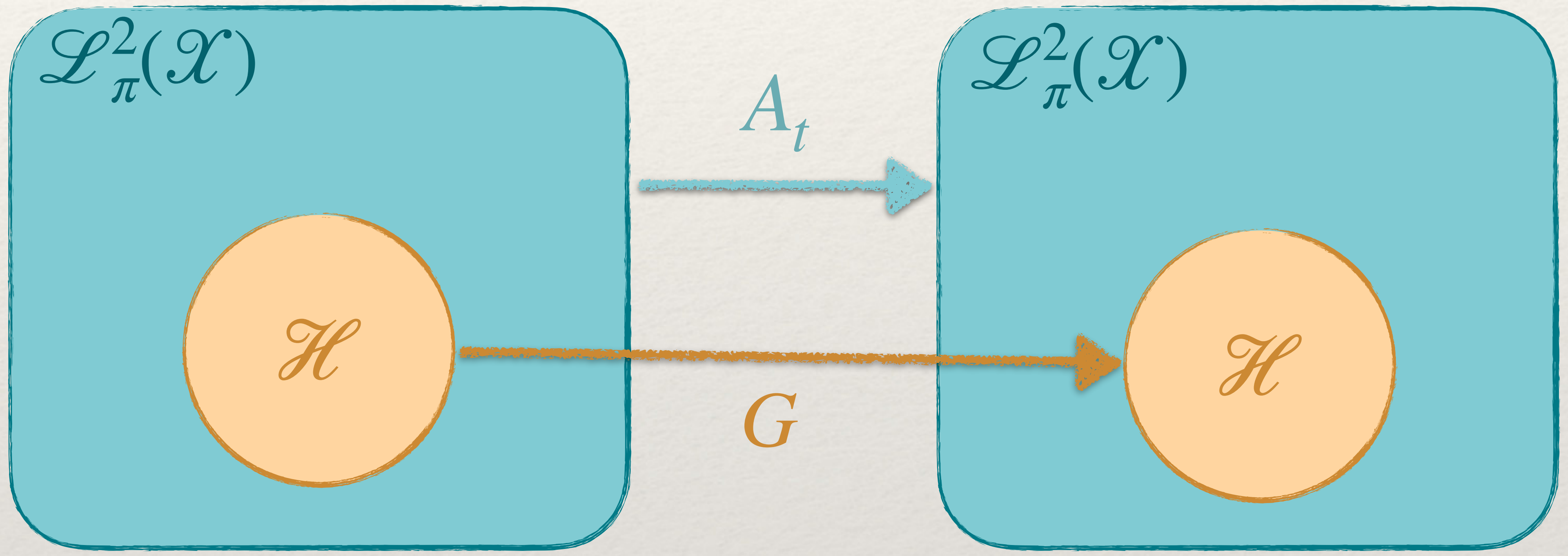
- Representation a priori chosen or learned
- Regression can be w.r.t. various losses and regularisation types
- Both can be w/o prior knowledge
- We might care of various tasks

Reminder on SLT of operator regression

Since we don't know $L^2_\pi(\mathcal{X})$ we restrict A_t to a chosen hypothesis space $\mathcal{H}$ and look for an operator $G : \mathcal{H} \to \mathcal{H}$ such that $A_t\langle w, \phi(\,\cdot\,)\rangle \approx \langle Gw, \phi(\,\cdot\,)\rangle$, leading to

Risk minimisation:

$$\boxed{\mathcal{R}(G) = \mathbb{E}_{X_s \sim \pi} \|\phi(X_{s+t}) - G^*\phi(X_s)\|^2}$$



$$Gh_i = \lambda_i h_i \;\Rightarrow\; \|(\lambda_i\, I - A_t)^{-1}\|^{-1} \le \|A_t h_i - \lambda_i h_i\|_{L^2_\pi(\mathcal{X})} \le \underbrace{\|A_t|_{\mathcal{H}} - G\|_{\mathcal{H}\to L^2_\pi(\mathcal{X})}}_{\mathcal{E}(G)} \underbrace{\frac{\|h_i\|_{\mathcal{H}}}{\|h_i\|_{L^2_\pi(\mathcal{X})}}}$$

$$\|\mathbb{E}[h(X_{s+t}) \,|\, X_s = \cdot\,] - Gh\| \le \mathcal{E}(G)\|h\|_{\mathcal{H}}$$

t-step ahead prediction

operator norm error (excess risk)

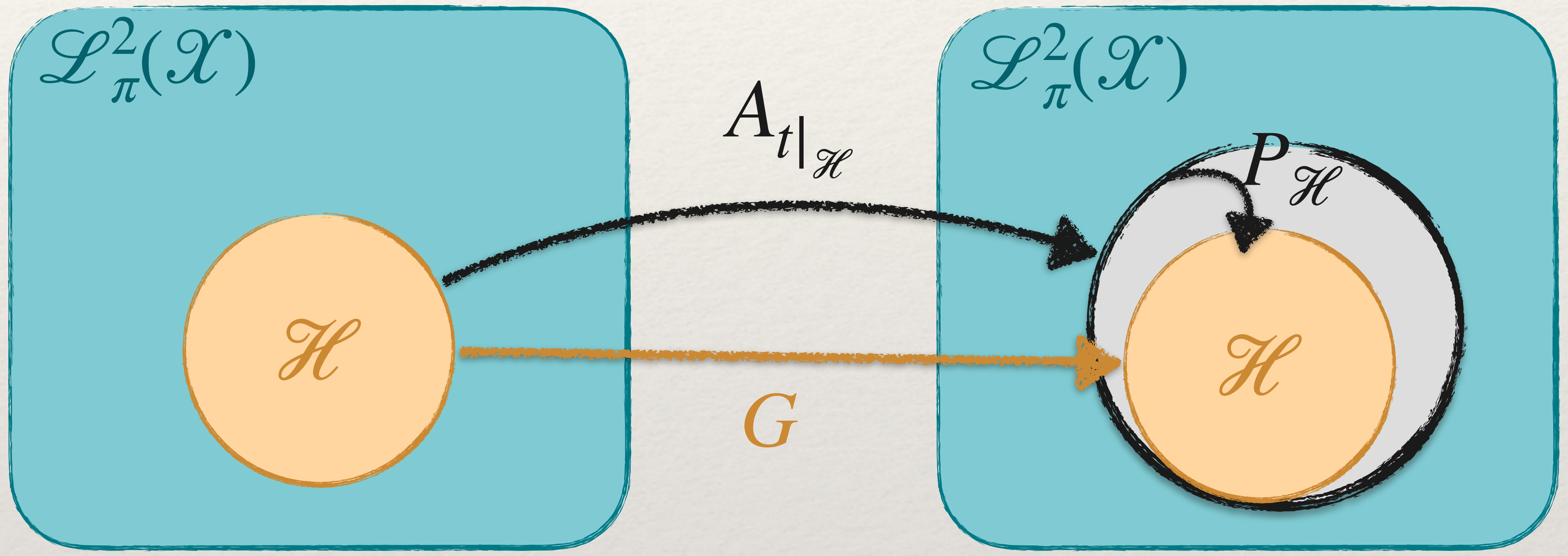
metric distorsion

Reminder on SLT of operator regression

Since we don't know $L^2_\pi(\mathscr{X})$ we restrict A_t to a chosen hypothesis space $\mathscr{H}$ and look for an operator $G : \mathscr{H} \to \mathscr{H}$ such that $A_t\langle w, \phi(\cdot)\rangle \approx \langle Gw, \phi(\cdot)\rangle$, leading to

Metric distortion via covariance operator:

$$\boxed{\eta(h) = \frac{\|h\|_{\mathscr{H}}}{\|C^{1/2}h\|_{\mathscr{H}}} \qquad C = \mathbb{E}_{X\sim\pi}\,\phi(X)\otimes\phi(X)}$$



Projection operator: $P_{\mathscr{H}} f = \operatorname*{argmin}_{h\in\mathscr{H}} \|f-h\|_{\mathscr{L}^2_\pi},\ f\in\mathscr{L}^2_\pi$

Estimation error decomposition

$$\boxed{\mathscr{E}(\hat{G}) \le \underbrace{\|(I-P_{\mathscr{H}})A_{t|_{\mathscr{H}}}\|_{\mathscr{H}\to\mathscr{L}^2_\pi}}_{\text{representation error}} + \underbrace{\|P_{\mathscr{H}}A_{t|_{\mathscr{H}}} - G\|_{\mathscr{H}\to\mathscr{L}^2_\pi}}_{\text{estimator's bias}} + \underbrace{\|G-\hat{G}\|_{\mathscr{H}\to\mathscr{L}^2_\pi}}_{\text{estimator's variance}}}$$

$\hat{G}$ is empirical version of G

What is the optimal representation?

Typically we have two situations, $\mathscr{H}$ is either finite or infinite-dimensional RKHS

- RKHS is a span of dictionary of functions, i.e. $\mathscr{H} = \mathrm{span}(z_j)_{j \in [d]} \subset \mathscr{L}^2_\pi(\mathscr{X})$
  - Representation error is controlled by letting $d \to \infty$
  - Without the prior knowledge, the representation error is a bottleneck

- RKHS $\mathscr{H}$ is given by some universal reproducing kernel $k\colon \mathscr{X} \times \mathscr{X} \to \mathbb{R}$
  - No representation error, i.e. $\|(I - P_{\mathscr{H}})A_{t|_{\mathscr{H}}}\|_{\mathscr{H} \to \mathscr{L}^2_\pi} = 0$
  - Learning guarantees depend on the effective dimension of $\mathscr{H}$ in $\mathscr{L}^2_\pi(\mathscr{X})$, and the regularity of A_t w.r.t. $\mathscr{H}$ (*the devil is in the tail eiegenvectors of covariance*)

What is the optimal representation?

Typically we have two situations, $\mathscr{H}$ is either finite or infinite-dimensional RKHS

- RKHS is a span of dictionary of functions, i.e. $\mathscr{H} = \mathrm{span}(z_j)_{j\in[d]} \subset \mathscr{L}^2_\pi(\mathscr{X})$
  - Representation error is controlled by letting $d \to \infty$
  - Without the prior knowledge, the representation error is a bottleneck

- Representation desiderata:
  - control the representation error, i.e. $\|(I - P_{\mathscr{H}})A_{t|_{\mathscr{H}}}\|_{\mathscr{H}\to\mathscr{L}^2_\pi}$
  - approximate well the operator $P_{\mathscr{H}}A_t \approx A_t$
  - align the geometries of $\mathscr{H}$ and $\mathscr{L}^2_\pi(\mathscr{X})$, i.e. $C \approx I$

SVD

What is the optimal representation?

- When A_t is compact, the good choice for $\mathscr{H}$ is its leading left singular subspace
  - the representation error is in general controlled $\|(I-P_{\mathscr{H}})A_{t|_{\mathscr{H}}}\|_{\mathscr{H}\to\mathscr{L}^2_\pi} \leq \sigma_d$ and if $A_t^*A_t = A_tA_t^*$ it is not even present
  - we approximate well, since $P_{\mathscr{H}}A_t$ is the best rank-d approximation of A_t
  - the geometry of $\mathscr{H}$ and $\mathscr{L}^2_\pi(\mathscr{X})$ are the same since the orthonormality of the singular functions implies $C = I$
- The general problem is to learn the SVD of $E_{Y|X}\colon \mathscr{L}^2_{\mu_Y} \to \mathscr{L}^2_{\mu_X}$ having only the samples of $(X, Y) \sim \mu_{X,Y}$

$\boxed{\text{We can estimate } \langle f, E_{Y|X}g\rangle_{\mathscr{L}^2_{\mu_X}} = \mathbb{E}[f(X)g(Y)]\,!}$

Linear Algebra meets Neural Networks

- We can learn $E_{Y|X} = \sum_{i\in\mathbb{N}_0} \sigma_i\, u_i \otimes v_i$ with neural networks $(\sigma_i^\theta, u_i^\theta, v_i^\theta)_{i\in[d]}$ via different variational principles
  - Deep projections (*ICLR2024*):

$$\boxed{\|C_X^{\dagger/2} C_{XY} C_Y^{\dagger/2}\|_F^2 \geq \|C_{XY}\|_F^2/(\|C_X\|\,\|C_Y\|)}$$

$$\max_{(u_i,v_i)_{i\in[d]}} \|P_{\mathscr{H}_u} E_{Y|X} P_{\mathscr{H}_v}\|^2_{\mathrm{HS}(\mathscr{L}^2_{\mu_Y},\mathscr{L}^2_{\mu_X})} \qquad C_Y = \mathbb{E}_{(X,Y)}[u(X)v(Y)^\top]$$

  - Eckhart-Mirsky-Young (*NeurIPS2024*):

$$\boxed{\mathrm{tr}\big(\Sigma C_X \Sigma C_Y - 2\Sigma C_{XY}\big)}$$

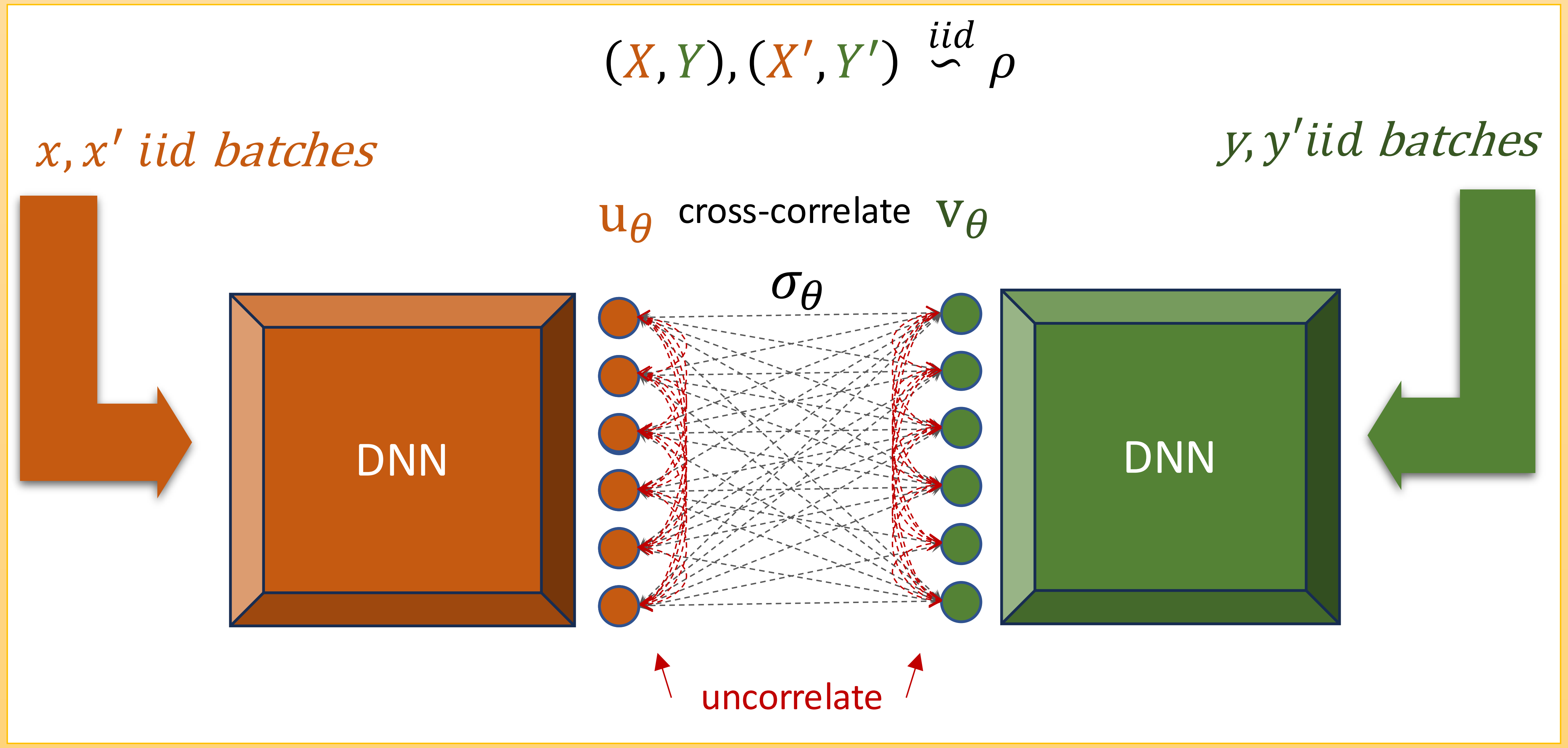
$$\min_{(\sigma_i,u_i,v_i)_{i\in[d]}} \|E_{Y|X} - \sum_{i\in[d]} \sigma_i\, u_i \otimes v_i\|^2_{\mathrm{HS}(\mathscr{L}^2_{\mu_Y},\mathscr{L}^2_{\mu_X})} - \|E_{Y|X}\|^2_{\mathrm{HS}(\mathscr{L}^2_{\mu_Y},\mathscr{L}^2_{\mu_X})}$$

subject to $C_X = \mathbb{E}_X[u(X)u(X)^\top] = I$ and $C_Y = \mathbb{E}_Y[v(X)v(X)^\top] = I$

Linear Algebra meets Neural Networks

$$\mathscr{L}_\gamma(\theta) := \mathbb{E}_{(X,Y),(X',Y')\sim\rho\; iid} L(u^\theta(X), u^\theta(X'), v^\theta(Y), v^\theta(Y'), \sigma^\theta) + \gamma\, R(u^\theta(X), u^\theta(X'), v^\theta(Y), v^\theta(Y'))$$

LEARING THE REPRESENTATION OF CONDITINAL PROBABILITY



Loss functional (cross-correlate):

$$L(u,u',v,v',s) = \frac{1}{2}\left(u^\top \mathrm{diag}(s) v'\right)^2 + \frac{1}{2}\left(u'^\top \mathrm{diag}(s) v\right)^2 - (u-u')^\top \mathrm{diag}(s)(v-v')$$

Orthogonality constraints (uncorrelate):

$$R(u,u',v,v') = (u^\top u')^2 - (u-u')^\top(u-u') + (v^\top v')^2 - (v-v')^\top(v-v') + 2d$$

Regression in the representation space via SVD:

$$(\hat{\mathbb{E}}[u^\theta \Sigma^\theta (u^\theta)^\top])^{-1/2} (\hat{\mathbb{E}}[u^\theta \Sigma^\theta (v^\theta)^\top]) (\hat{\mathbb{E}}[v^\theta \Sigma^\theta (v^\theta)^\top])^{-1/2} = \hat{U}\hat{\Sigma}\hat{V}^\top$$

$$\sigma^\theta \leftarrow \hat{\sigma} \qquad u^\theta \leftarrow \hat{U}^\top (\Sigma^\theta)^{1/2} u^\theta \qquad v^\theta \leftarrow \hat{V}^\top (\Sigma^\theta)^{1/2} v^\theta$$

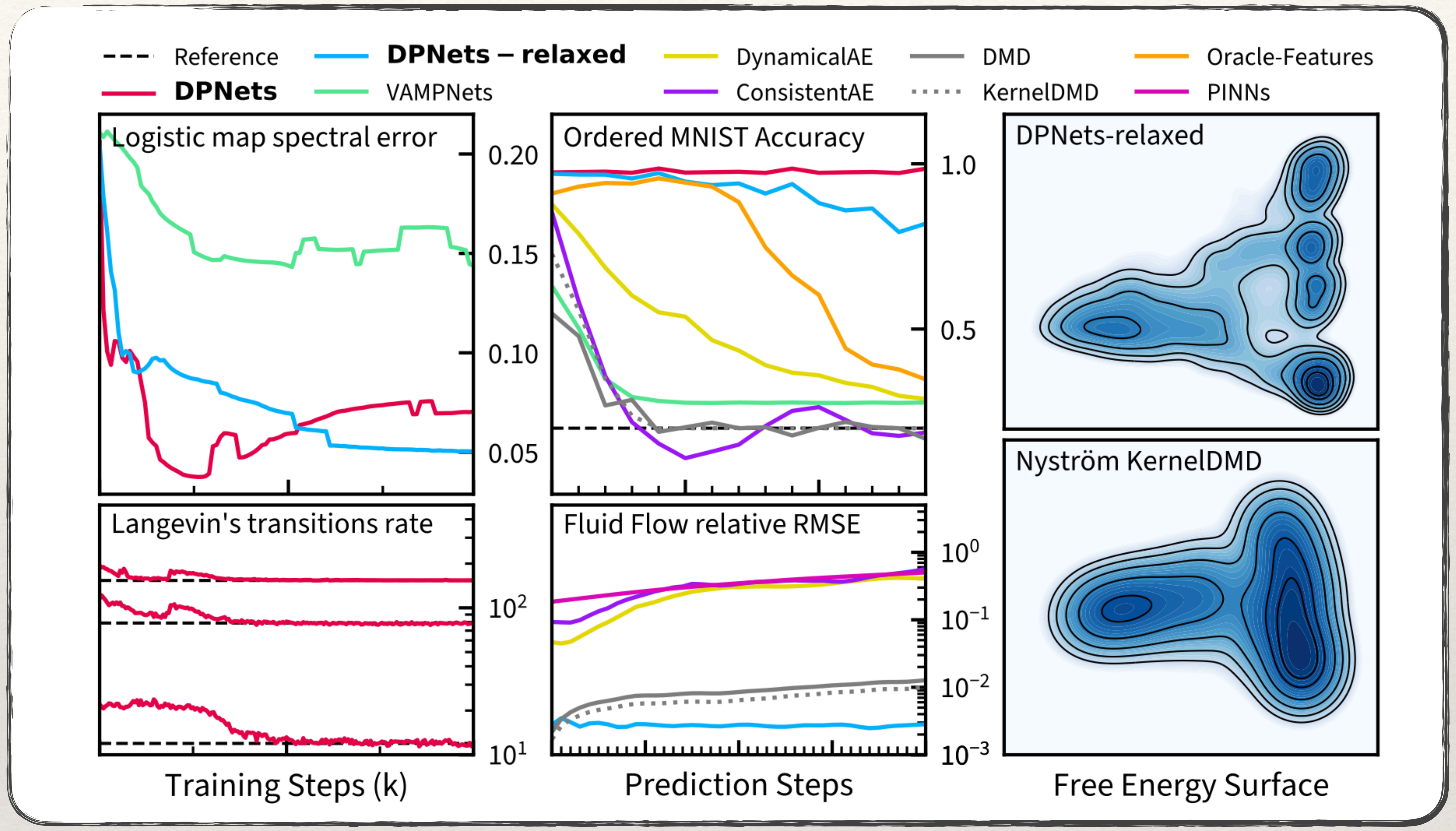
Learned truncated SVD

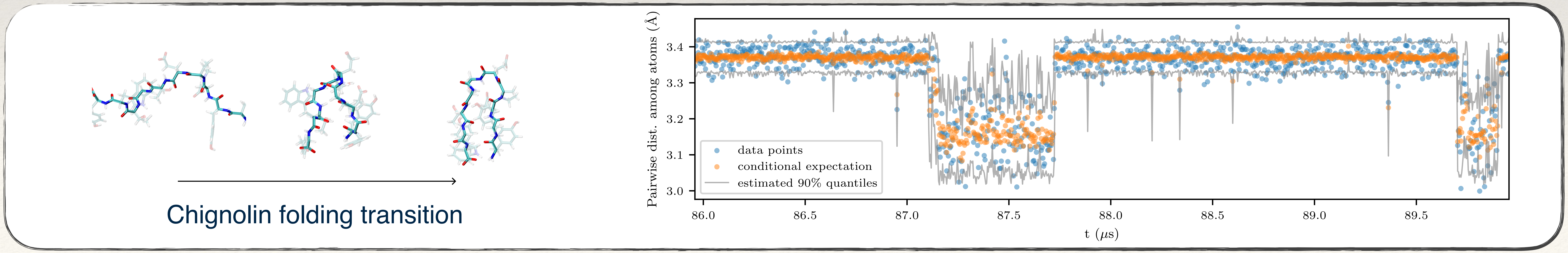
$$\widehat{E}_{Y|X} = \sum_{j=1}^{d} \hat{\sigma}_j^\theta\, \hat{u}_j^\theta \otimes \hat{v}_j^\theta$$

Back to the example of digits…

Other examples…

- Noisy Logistic map
- 1D triple well potential Langevin dynamics
- Fluid flow around cylinder
- Folding of a mini-protein in water





What about theory?

Key advantages of representation learning + regression:

(1) It extracts statistics directly from the trained operator without retraining or resampling

(2) We get best of both worlds kernel methods (strong statistical theory) and DL (representation power of NN architectures)

$$\widehat{\mathbb{P}}_{\theta}[Y \in B \mid X \in A] - \mathbb{P}[Y \in B \mid X \in A] = O_{\mathbb{P}}\left(\frac{1}{\sqrt{n}} + \overbrace{\sqrt{\frac{\mathbb{P}[Y \in B]}{\mathbb{P}[X \in A]}}}^{\text{Fundamental Statistical Limit}} \left(\overbrace{\sigma^{\star}_{d+1}}^{\text{Bias}} + \underbrace{\mathcal{E}_{\theta}}_{\text{Optimisation Error}} + \overbrace{\sqrt{d/n}}^{\text{Statistical Error}} \right) \right)$$

Physics-informed learning with the generator

- Family of TOs $A_t : \mathscr{L}^2_\pi(\mathscr{X}) \to \mathscr{L}^2_\pi(\mathscr{X})$, $t \geq 0$, forms a continuous semigroup characterised by the **infinitesimal generator (IG)** $L = \lim_{t\to 0^+} (A_t - I)/t : \mathscr{L}^2_\pi(\mathscr{X}) \to \mathscr{L}^2_\pi(\mathscr{X})$, an **unbounded operator** with $\mathrm{dom}(L) = \left\{ f \in \mathscr{L}^2_\pi \mid \sum_{i\in[d]} \|\partial_i f\|^2_{\mathscr{L}^2_\pi} < \infty \right\}$ given by

$$(Lf)(x) = \nabla f(x)^\top a(x) + \frac{1}{2}\mathrm{Tr}\left[b(x)^\top (\nabla^2 f(x)) b(x)\right], \quad \forall f \in \mathrm{dom}(L)$$

- When the process is additionally time reversal invariant, IG is **self-adjoint** operator that introduces kinetic energy kernel, which often can be written in the **Dirichlet form** $s : \mathbb{R}^d \to \mathbb{R}^p$

$$\mathfrak{E}_\pi[f,g] = -\langle f, Lg \rangle_{\mathscr{L}^2_\pi} = \int_{\mathscr{X}} \nabla f(x)^\top s(x) s(x)^\top \nabla g(x) \pi(dx) \qquad \mathfrak{E}_{X\sim\pi} f(X) = \mathbb{E}_{X\sim\pi} \|s(X)^\top \nabla f(X)\|^2$$

- Solving an SDE: from IG to TO and back with IG's exponential and **resolvent operator**, both **bounded** operators

$$A_t = e^{tL} \qquad R_\mu = (\mu I - L)^{-1} = \int_0^{+\infty} A_t e^{-\mu t}\, dt, \quad \mu > 0$$

- **Spectral decomposition** of IG allows one to efficiently handle both, that is $L = \sum_{i=0}^{\infty} \lambda_i f_i \otimes f_i$ implies

$$(\mu I - L)^{-1} = \sum_{i=0}^{\infty} \overbrace{(\mu - \lambda_i)^{-1}}^{\nu_i} f_i \otimes f_i \qquad \mathbb{E}[f(X_t) \mid X_0 = x] = \sum_{i\in\mathbb{N}} e^{\lambda_i t} \langle f, f_i \rangle_{\mathscr{L}^2_\pi} f_i(x) \quad (\forall f)(\forall x)(\forall t)$$

- Hence, to build kinetic models we need to **learn leading eigenpairs of IG**. Since the obvious choice of Galerkin projections suffers from **spurious spectral estimation** due to unbounded nature of L, we approach **the problem through the resolvent**.

Physics-informed learning with the generator

- When estimating the largest eigenvalues of the resolvent $R_\mu f_i = \nu_i f_i$, the quality of estimator's decomposition $\hat{G}\hat{h}_i = \hat{\nu}_i \hat{h}_i$ is determined by the **alignment of norms** in the domain $\mathscr{W} = \{f \in \mathrm{dom}(L) \mid \|f\|_{\mathscr{W}} < \infty\}$ and $\mathscr{H}$ and the **estimation error.**

$$|\nu_i - \hat{\nu}_i| \leq \mathscr{E}(\hat{G})\ \eta(\hat{h}_i)$$

Metric distortion: $\|\hat{h}_i\|_{\mathscr{H}} / \|\hat{h}_i\|_{\mathscr{W}}$

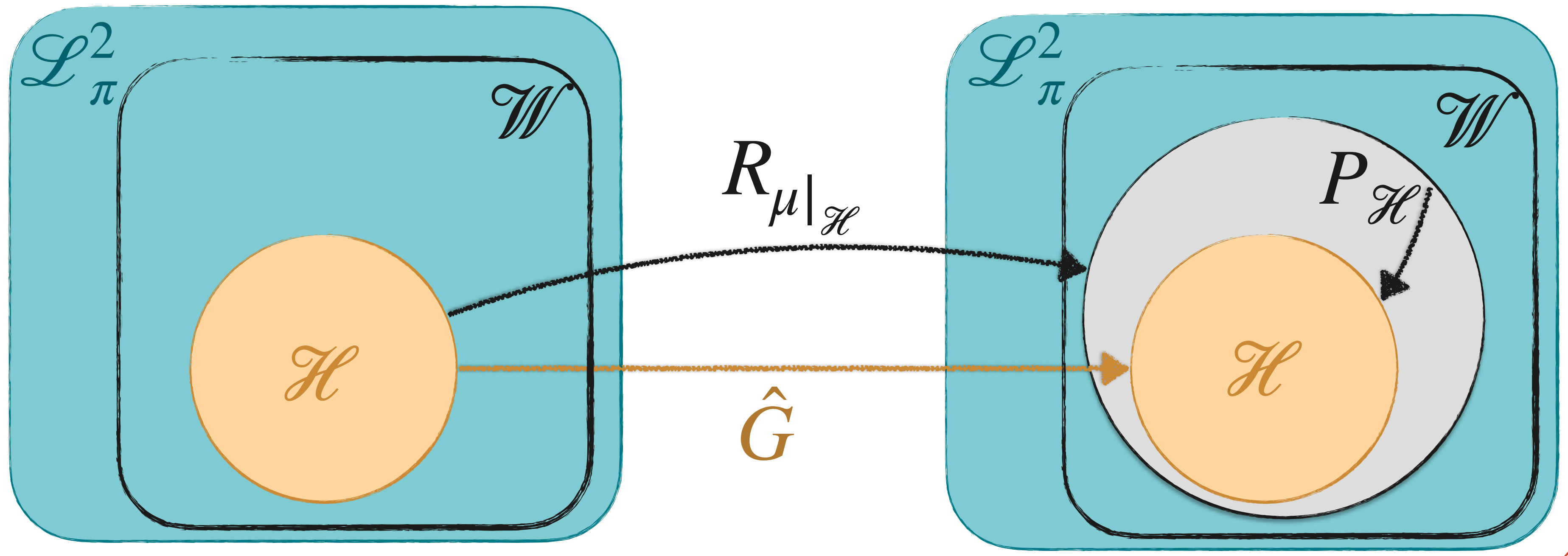
Estimation error:

Representation error

Estimator's error

$$\mathscr{E}(\hat{G}) = \|R_{\mu|_{\mathscr{H}}} - \hat{G}\|_{\mathscr{H}\to\mathscr{W}} \leq \|(I - P_{\mathscr{H}})A_{\pi|_{\mathscr{H}}}\|_{\mathscr{H}\to\mathscr{W}} + \|P_{\mathscr{H}} R_{\mu|_{\mathscr{H}}} - \hat{G}\|_{\mathscr{H}\to\mathscr{W}}$$

Projection operator: $P_{\mathscr{H}} f = \mathrm{argmin}_{h \in \mathscr{H}} \|f - h\|_{\mathscr{W}},\ f \in \mathrm{dom}(L)$



What is the good choice of geometry to make efficient and reliable algorithms ?

Physics-informed learning with the generator

- When estimating the largest eigenvalues of the resolvent $R_\mu f_i = \nu_i f_i$, the quality of estimator's decomposition $\hat{G}\hat{h}_i = \hat{\nu}_i \hat{h}_i$ is determined by the **alignment of norms** in the domain $\mathscr{W} = \{f \in \mathrm{dom}(L) \mid \|f\|_{\mathscr{W}} < \infty\}$ and $\mathscr{H}$ and the **estimation error.**

$$|\nu_i - \hat{\nu}_i| \leq \mathscr{E}(\hat{G})\ \eta(\hat{h}_i)$$

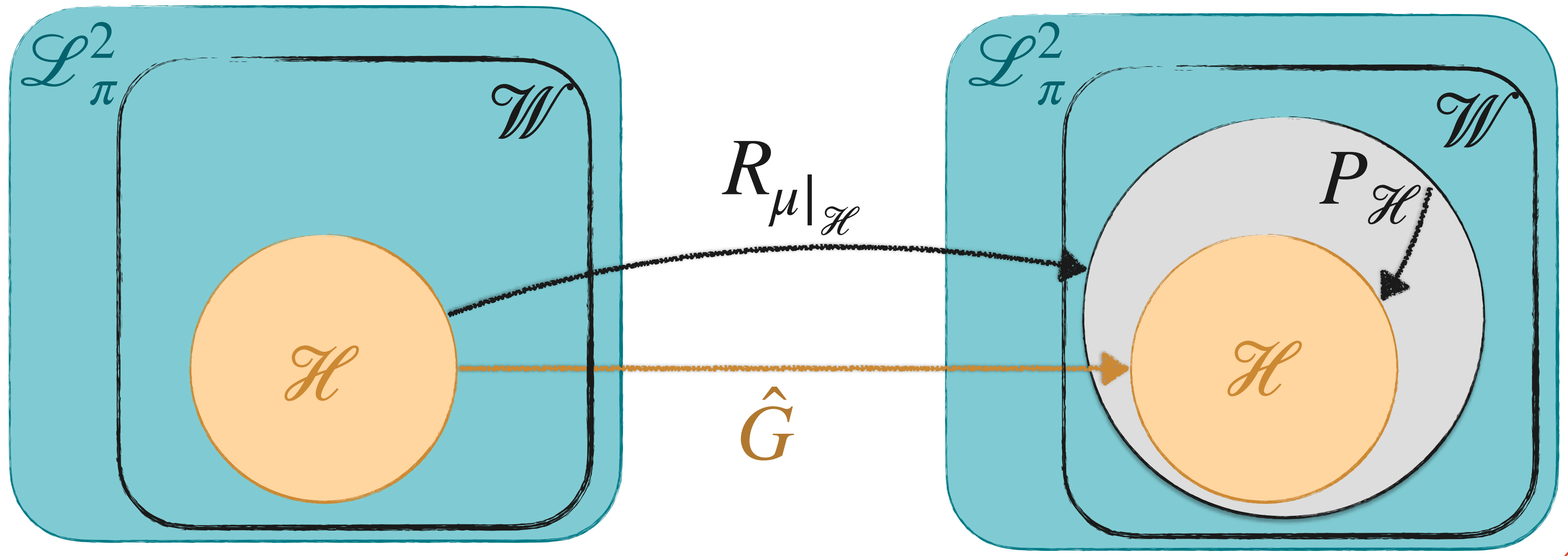
→ **Metric distortion:** $\|\hat{h}_i\|_{\mathscr{H}} \,/\, \|\hat{h}_i\|_{\mathscr{W}}$

Estimation error:

Representation error | Estimator's error

$$\mathscr{E}(\hat{G}) = \|R_{\mu|_{\mathscr{H}}} - \hat{G}\|_{\mathscr{H}\to\mathscr{W}} \leq \|(I-P_{\mathscr{H}})A_{\pi|_{\mathscr{H}}}\|_{\mathscr{H}\to\mathscr{W}} + \|P_{\mathscr{H}}R_{\mu|_{\mathscr{H}}} - \hat{G}\|_{\mathscr{H}\to\mathscr{W}}$$

Projection operator: $P_{\mathscr{H}} f = \mathrm{argmin}_{h\in\mathscr{H}} \|f-h\|_{\mathscr{W}},\ f \in \mathrm{dom}(L)$



- Since R_μ is bounded we can learn it via regression in RKHS, however computing its action by inverting, i.e. integral transform is not feasible! So, we **fight fire with fire** by adapting $\mathscr{W}$

$$\|f\|^2_{\mathscr{W}} = \langle f, (\mu I - L)f\rangle_{\mathscr{L}^2_\pi} = \mathbb{E}_{X\sim\pi}[\mu\,|f(x)|^2 + \|s(x)^\top \nabla f(x)\|^2] =: \mathfrak{E}^\mu_{X\sim\pi} f(X)$$

- Chosen geometry of $\mathrm{dom}(L)$ leads to the notion of **energy based risk functional**

$$\mathscr{R}(G) = \mathfrak{E}^\mu_{X\sim\pi}\,\|\chi_\mu(X) - G^*\phi(X)\|^2_{\mathscr{H}} = \|R_\mu - G\|^2_{\mathrm{HS}(\mathscr{H},\mathscr{W})}$$

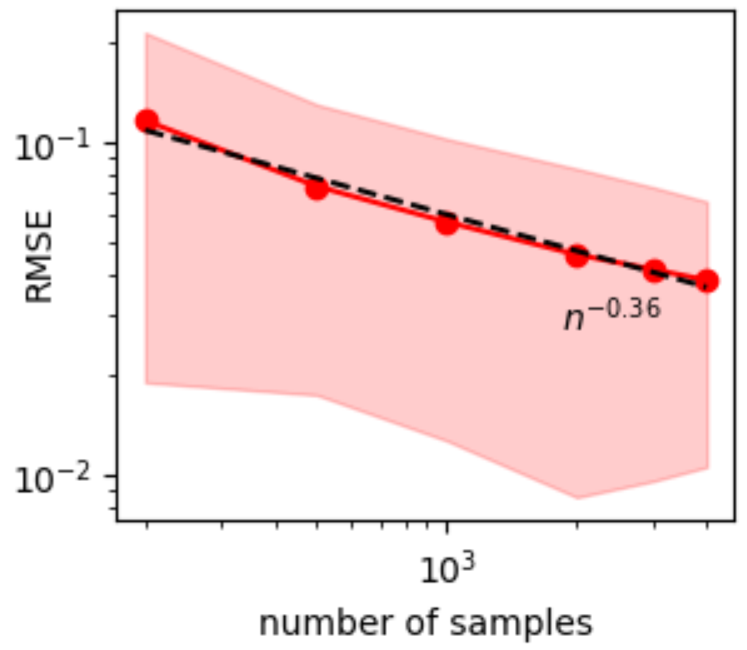
that balances the inverse $\mathscr{R}(G) = \|R_\mu^{1/2} - R_\mu^{-1/2}G\|^2_{\mathrm{HS}(\mathscr{H},\mathscr{L}^2_\pi)}$, and can be efficiently empirically minimised in closed form

Physics-informed learning with the generator

- Summary of guarantees for RRR with **universal bounded kernel** compared to SOTA

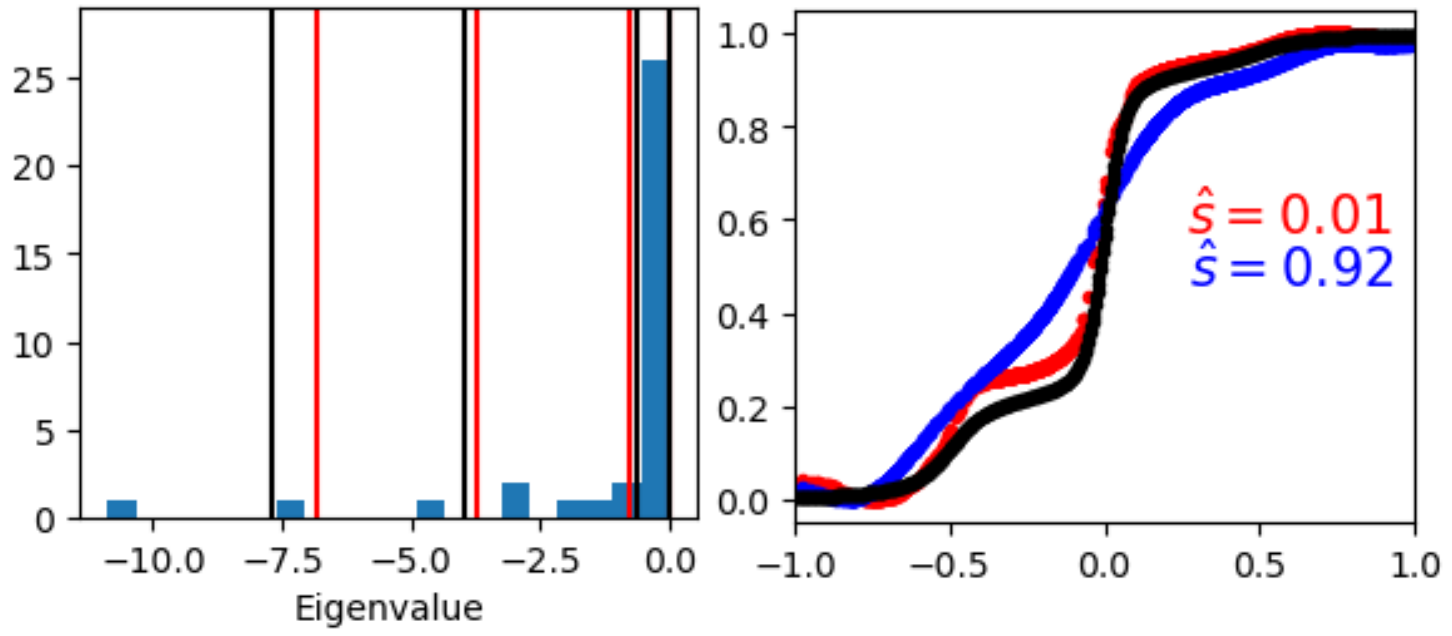
Aspect	Cabbanes & Bach 2024	Hou et al. 2024	Pillaud-Vivien & Bach 2023	Our work
Covers many SDEs	✗ (only Laplacian)	✓	✗ (only Langevin)	✓
Risk metric	$\mathcal{L}^2_\pi(\mathcal{X})$ metric	$\mathcal{L}^2_\pi(\mathcal{X})$ metric	$\mathcal{L}^2_\pi(\mathcal{X})$ metric	energ
Physics-informed method	✗	✓(full info. needed)	✗	✓ (partial info. needed)
Avoids spurious eigenvalues	✗	✗	✗	✓
IG error bound	$\mathcal{O}(n^{-\frac{d}{2(d+1)}})$	$\mathrm{Var} = \mathcal{O}(\frac{d^2}{\gamma^2\sqrt{n}})$	$\mathcal{O}(n^{-\frac{1}{4}})$	$\mathcal{O}(n^{-\frac{\alpha}{2(\alpha+\beta)}}),\ \alpha \geq \tau$ $\mathcal{O}(n^{-\frac{\alpha}{2(\beta+\tau)}}),\ \alpha < \tau$
Spectral rates	✗	✗	✗	✓
Time complexity	$\mathcal{O}(n^2 + n^{3/2}d)$	$\mathcal{O}(n^3d^3)$	$\mathcal{O}(n^3d^3)$	$\mathcal{O}(r\,n^2d^2)$

Cox-Ingersoll-Ross



learning rates

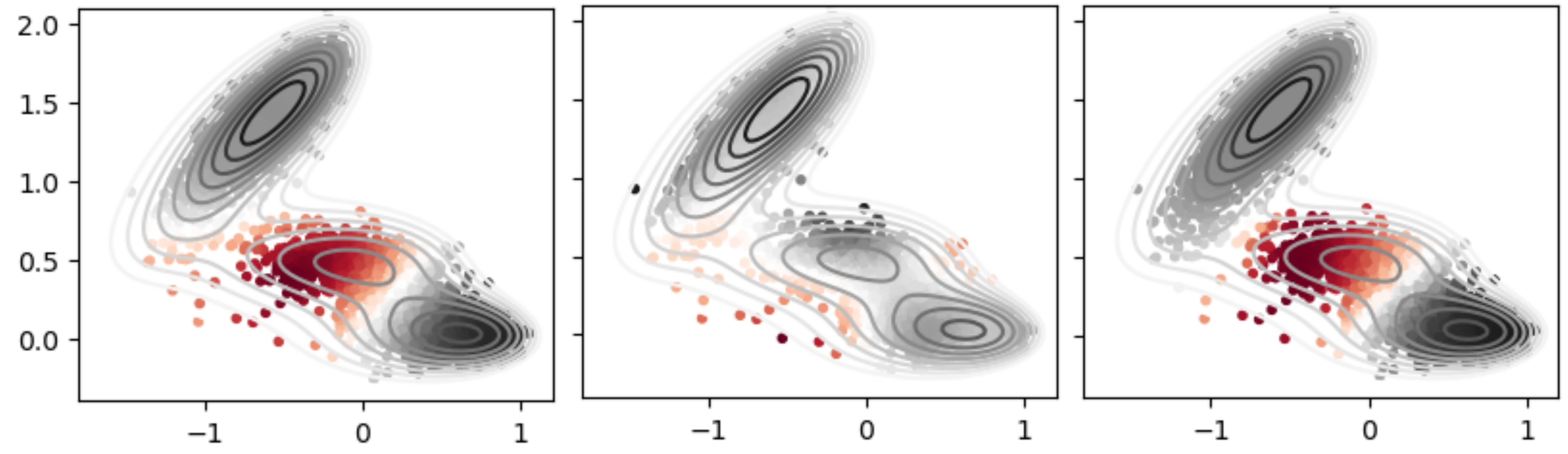
1D Langevin SDE



non-spurious spectra

eigenfunction estimation

2D Langevin

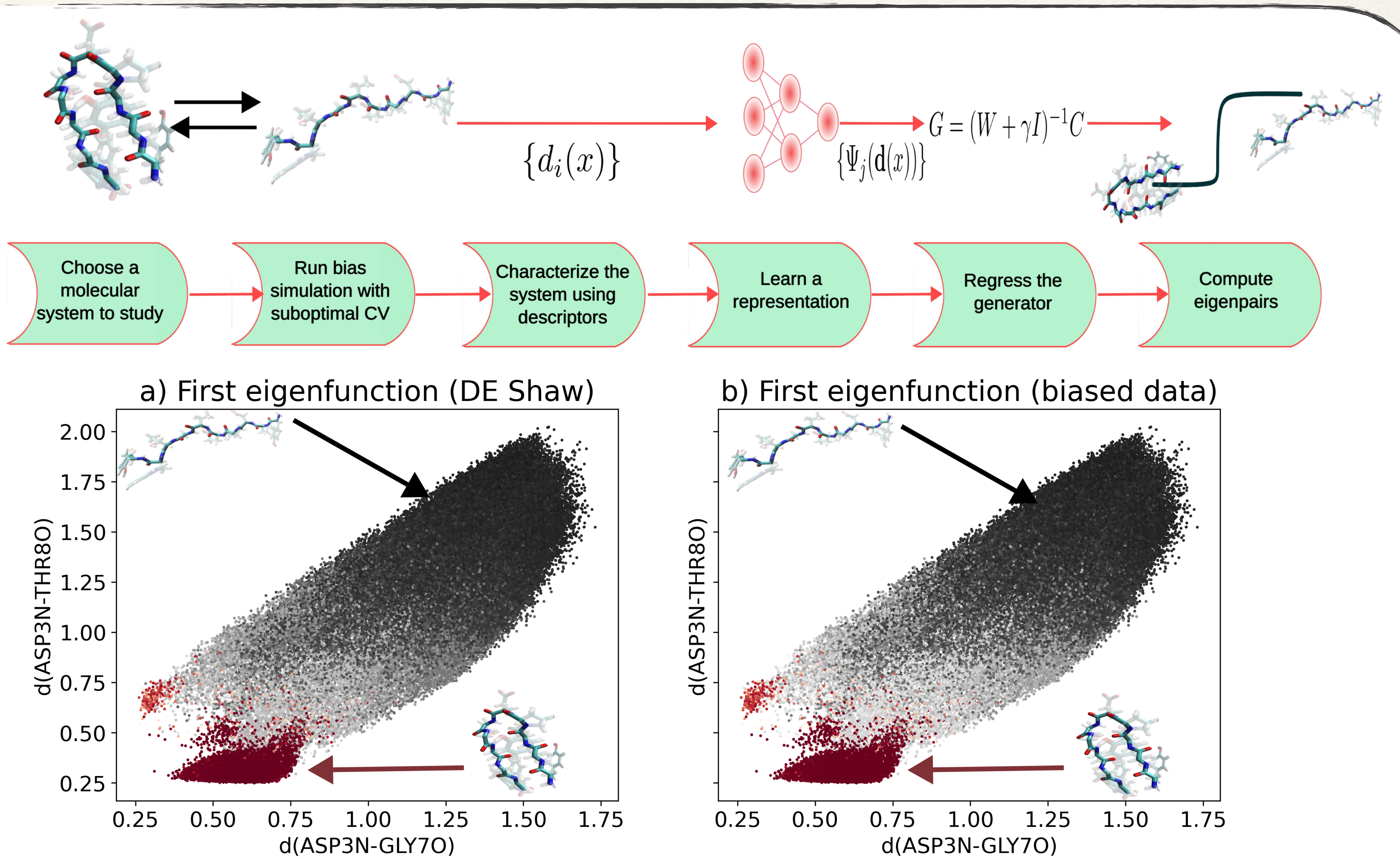


IG vs TO sample efficiency

Physics-informed learning with the generator

- PI representation learning (time-reversal invariant process in equilibrium)

  - EMY principle w.r.t. energy norm
  - Learns kinetic model from static data
  - Neatly combined with enhanced sampling (*control the process to discover meta-stable states*)



We tested our method on a **1** μs long biased simulation and compared it to the results obtained when training on a **107** μs unbiased simulation

References and Code

- V. Kostic, P. Novelli, A. Maurer, C. Ciliberto, L. Rosasco, M. Pontil. Learning dynamical systems via Koopman operator regression in reproducing kernel hilbert spaces. NeurIPS 2022.

- V. Kostic, K. Lounici, P. Novelli, M. Pontil. Koopman operator learning: sharp spectral rates and spurious eigenvalues. NeurIPS 2023.

- G. Meanti, A. Chatalic, V. Kostic, P. Novelli, M. Pontil, L. Rosasco. Estimating Koopman operators with sketching to provably learn large scale dynamical systems. NeurIPS 2023.

- V. Kostic, P. Novelli, R. Grazzi, K. Lounici, M. Pontil. Learning invariant representations of time-homogeneous stochastic dynamical systems. ICLR 2024.

- V. Kostic, K. Lounici, P. Inzerilli, P. Novelli., M. Pontil. Consistent long-term forecasting of ergodic dynamical systems. ICML 2024.

- K. Lounici, V Kostic, G. Pacreau, G. Turri, P. Novelli, M. Pontil Neural Conditional Probability for Statistical Inference, NeurIPS 2024.

- V. Kostic, K. Lounici, H. Halconruy, T. Devergne, M. Pontil. Learning the infinitesimal generator of stochastic diffusion processes, NeurIPS 2024

- T. Devergne, V. Kostic, M. Parrinello, M. Pontil. From biased to unbiased dynamics: an infinitesimal generator approach. NeurIPS 2024

- V. Kostic, K. Lounici, H. Halconruy, T. Devergne, P. Novelli, M. Pontil. Learning the infinitesimal generator of stochastic diffusion processes, NeurIPS 2024

Code: https://github.com/Machine-Learning-Dynamical-Systems/kooplearn