Structure-Preserving Learning of Hamiltonian Systems

Juan-Pablo Ortega
(joint with Jianyu Hu and Daiying Yin)

Nanyang Technological University, Singapore

Trends in Mathematical Sciences. Friedrich-Alexander Universität Erlangen, June 2024.

Contents

Context and objectives

Hamiltonian systems (in Darboux coordinates)

$$\dot{\mathbf{z}}(t) = X_H(\mathbf{z}(t)) := J\nabla H(\mathbf{z}(t)), \quad t \in [0, T],$$

where $\mathbf{z} = (\mathbf{q}, \mathbf{p}) \in \mathbb{R}^{2d}$ is the position and momentum vector,

- $$J = \begin{pmatrix} 0 & \mathbb{I}_d \\ -\mathbb{I}_d & 0 \end{pmatrix}$$ is the canonical symplectic matrix.
- $H : \mathbb{R}^{2d} \longrightarrow \mathbb{R}$ is a Hamiltonian function.
- Hamilton's equations

$$\dot{q}^i = \frac{\partial H}{\partial q_i}, \quad \dot{p}_i = \frac{\partial H}{\partial p^i}, \quad i = 1, \ldots, d.$$

Designed for simple mechanical systems ($H = T + V$) and obtained out of a variational principle (Hamilton's principle).

Going beyond simple mechanical systems

Hamiltonian mechanics on symplectic manifolds

(M, ω) symplectic manifold $H : M \longrightarrow \mathbb{R}$ Hamiltonian function.

$$\mathbf{i}_{X_h}\omega = \mathbf{d}H$$

Examples: classical mechanics on non-Euclidean configuration spaces and Lie groups: pendula, robotic arms, rigid body mechanics, fluids.

Hamiltonian mechanics on Poisson manifolds

A Poisson manifold $(P, \{\cdot, \cdot\})$. $\{,\} : C^\infty(P) \times C^\infty(P) \to C^\infty(P)$ is a bilinear operation such that:

(i) $(C^\infty(P), \{,\})$ is a Lie algebra.

(ii) $\{,\}$ is a derivation in each factor, that is,

$$\{FG, H\} = \{F, H\}G + F\{G, H\}, \text{ for all } F, G, \text{ and } H \in C^\infty(P).$$

Poisson mechanics examples

Hamiltonian vector field: $X_H[F] = \{F, H\}, \quad \text{for all } F \in C^\infty(P)$.

- Symplectic case: $\{F, G\}(z) = \omega(X_F(z), X_G(z))$.
- Lie-Poisson mechanics on duals $\mathfrak{g}^*$ of Lie algebras:

$$\{F, G\}_\pm(\mu) = \pm \left\langle \mu, \left[\frac{\delta F}{\delta \mu}, \frac{\delta G}{\delta \mu}\right] \right\rangle, \ \mu \in \mathfrak{g}^* \text{ and } F, G \in C^\infty\left(\mathfrak{g}^*\right)$$

$$X_H(\mu) = \mp \mathrm{ad}^*_{\frac{\delta H}{\delta \mu}} \mu, \ \ \mu \in \mathfrak{g}^*.$$

A short description of some physical problems that can be written in Lie–Poisson form and related Poisson brackets.

Problem	Reference
Rigid body	Holm, Schmah, and Stoica (2009) Marsden and Ratiu (2013)
Heavy top	Holm et al. (2009) Marsden and Ratiu (2013)
Underwater vehicles	Leonard (1997) Leonard and Marsden (1997) Holmes, Jenkins, and Leonard (1998)
Plasmas	Morrison (1980), Marsden and Weinstein (1982), Holm, Marsden, Ratiu, and Weinstein (1985), Holm and Tronci (2010)
Fluids	Marsden and Weinstein (1983), Marsden, Ratiu, and Weinstein (1984), Holm et al. (1985), Morrison (1998), Morrison, Francoise, Naber, and Tsou (2006),
Geophysical fluid dynamics	Weinstein (1983), Holm (1986), Salmon (2004)
Complex and nematic fluids	Holm (2002), Gay-Balmaz and Ratiu (2009), Gay-Balmaz and Tronci (2010)
Molecular strand dynamics	Ellis, Gay-Balmaz, Holm, Putkaradze, and Ratiu (2010), Gay-Balmaz, Holm, Putkaradze, and Ratiu (2012)
Fluid–structure interactions	Gay-Balmaz and Putkaradze (2019)
Hybrid quantum–classical dynamics	Gay-Balmaz and Tronci (2022), Gay-Balmaz and Tronci (2023)

Taken from [EGBHP24]

Now the objective

Solve the inverse problem

- Find the Hamiltonian
  - What Hamiltonian? Problem intrinsically ill-posed.
- Out of observations of
  - Noisy realizations of the Hamiltonian vector field.
  - Other options: discrete-time temporal traces: implies learning a structure-preserving integrator. Choices involved.
  - Assume access to full state-space observations.
  - Formulation of a global solution not using local coordinates. Compare with [JZKK22, EGBHP24].
- Using Reproducing Kernel Hilbert Spaces (RKHS): Why?
- Imposing structure preservation
  - The estimated system will be Hamiltonian despite the presence of approximation and estimation errors.

Observation data regime

The random samples consist of

$$\{\mathbf{Z}_N, \mathbf{X}_{\sigma^2,N}\} := \{(\mathbf{Z}^{(n)})_{n=1}^N, (\mathbf{X}_{\sigma^2}^{(n)})_{n=1}^N\} \xrightarrow{realization} \{\mathbf{z}_N, \mathbf{x}_{\sigma^2,N}\}.$$

- $\mathbf{Z}^{(n)}$ are the phase space vectors containing the positions and the momenta of the system and they are IID random variables with the same distribution μ_Z.
- The noisy vector fields $\mathbf{X}_{\sigma^2}^{(n)} = X_H(\mathbf{Z}^{(n)}) + \boldsymbol{\varepsilon}^{(n)}$ where $\boldsymbol{\varepsilon}^{(n)}$ are IID random variables with mean zero and variance σ^2 and are independent to $\mathbf{Z}^{(n)}$.

Machine learning methods

First approach: kernel ridge regression, Hamiltonian and Lagrangian neural networks.

Construct an empirical quadratic risk functional

$$\frac{1}{N}\sum_{n=1}^{N}\|\mathbf{f}(\mathbf{Z}^{(n)})-\mathbf{X}_{\sigma^2}^{(n)}\|^2, \tag{1.1}$$

and find the least squares (or ridge) estimator of the vector field $\mathbf{f}$ over a hypothesis function space, such as RKHS or neural network classes.

- Not structure-preserving: no guarantee that the learned vector field $\widehat{\mathbf{f}}$ is Hamiltonian.
- For some methods: Lack of error analysis. Non-convex optimization problems.

Structure-preserving kernel regression

Structure-preserving kernel regression: We search the vector field $\mathbf{f}$ with specific form $\mathbf{f} = \mathbf{f}_h := X_h$, where h is in the reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$ with kernel K.

Optimization problem: We consider the following optimization using the regularized empirical risk

$$\widehat{h}_{\lambda,N} := \underset{h \in \mathcal{H}_K}{\arg\min}\, \widehat{R}_{\lambda,N}(h), \tag{2.1}$$

$$\widehat{R}_{\lambda,N}(h) := \frac{1}{N} \sum_{n=1}^{N} \| X_h(\mathbf{Z}^{(n)}) - \mathbf{X}_{\sigma^2}^{(n)} \|^2 + \lambda \|h\|_{\mathcal{H}_K}^2. \tag{2.2}$$

Need to address:

- The well-posedness of the optimization problem.
- The convergence analysis of the structure-preserving kernel estimator $\widehat{h}_{\lambda,N}$ to the real Hamiltonian H with respect to the RKHS norm.

Structure-preserving kernel regression

We also consider the optimization problem associated to the regularized statistical risk

$$h_\lambda^* := \underset{h \in \mathcal{H}_K}{\arg\min}\, R_\lambda(h), \tag{2.3}$$

$$R_\lambda(h) := \|X_h - X_H\|^2_{L^2(\mu_Z)} + \lambda \|h\|^2_{\mathcal{H}_K} + \sigma^2. \tag{2.4}$$

Consistence: The regularized empirical and statistical risks are consistent within the RKHS in the sense that for every $h \in \mathcal{H}_K$, we have that

$$\lim_{N \to \infty} \mathbb{E}_{\boldsymbol{\varepsilon}} \left[\widehat{R}_{\lambda, N}(h) \right] = R_\lambda(h), \quad a.s.$$

RKHS: A crash course

A Mercer kernel on $\mathcal{X}$ is a positive-semidefinite symmetric function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Positive-semidefinite means that Gram matrices

$$G := [K(x_i, x_j)]_{i,j=1}^{n}$$

are positive semi-definite for any $x_1, \cdots, x_n \in \mathcal{X}$ and any given n.

Definition (RKHS)

Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a Mercer kernel on a nonempty set $\mathcal{X} \subseteq \mathbb{R}^d$. A Hilbert space $\mathcal{H}_K$ of real-valued functions on $\mathcal{X}$ endowed with the pointwise sum and pointwise scalar multiplication, and with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_K}$ is a reproducing kernel Hilbert space (RKHS) associated to K if:

(i) For all $x \in \mathcal{X}$, the function $K(x, \cdot) =: K_x \in \mathcal{H}_K$.

(ii) For all $x \in \mathcal{X}$ and for all $f \in \mathcal{H}_K$, the following reproducing property holds

$$f(x) = \langle f, K(x, \cdot) \rangle_{\mathcal{H}_K}.$$

Properties of RKHS

- There is a bijection between RKHSs and Mercer kernels.
- Given a kernel K, the corresponding RKHS $\mathcal{H}_K$ can be constructed as the completion of the span of elements of the form

$$f = \sum_{i=1}^{N} c_i K(x_i, \cdot), \;\; c_i \in \mathbb{R}, x_i \subset \mathcal{X}.$$

- Universal kernels: the Gaussian kernel on Euclidean spaces.

$$\mathcal{H}_K(\mathcal{Z}) = \overline{\operatorname{span}\left\{K_z \mid z \in \mathcal{Z}\right\}}.$$

Denote now by $\overline{\mathcal{H}_K(\mathcal{Z})}$ the uniform closure of $\mathcal{H}_K(\mathcal{Z})$. A kernel K is called universal if for any compact subset $\mathcal{Z} \subset \mathcal{X}$, we have that $\overline{\mathcal{H}_K(\mathcal{Z})} = C(\mathcal{Z})$.

Differential reproducing property

Theorem

Let $s \in \mathbb{N}$, and $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a Mercer kernel such that $K \in C_b^{2s+1}(\mathbb{R}^d \times \mathbb{R}^d)$. Then:

(i) *For any $x \in \mathbb{R}^d$ and $\alpha \in I_s$, $(D^\alpha K)_x \in \mathcal{H}_K$.*

(ii) *A differential reproducing property holds true for $\alpha \in I_s$:*

$$D^\alpha f(x) = \langle (D^\alpha K)_x, f \rangle_{\mathcal{H}_K} \quad \forall x \in \mathbb{R}^d, f \in \mathcal{H}_K. \tag{2.5}$$

(iii) *Denote $\kappa^2 = \|K\|_{C_b^{2s}(\mathbb{R}^d \times \mathbb{R}^d)}$. The inclusion $J : \mathcal{H}_K \hookrightarrow C_b^s(\mathbb{R}^d)$ is well-defined and bounded:*

$$\|f\|_{C_b^s} \leqslant \kappa \|f\|_{\mathcal{H}_K} \quad \forall f \in \mathcal{H}_K.$$

Operator representation

We define the operator A as

$$Ah = X_h, \quad h \in \mathcal{H}_K.$$

If $K \in C_b^3(\mathbb{R}^{2d} \times \mathbb{R}^{2d})$, the operator $A : \mathcal{H}_K \to L^2(\mathbb{R}^{2d}; \mu_Z; \mathbb{R}^{2d})$ is bounded linear. The adjoint operator A^* is

$$A^* g = \int_{\mathbb{R}^{2d}} g^T(x) J \nabla_1 K(x, \cdot) \, \mathrm{d}\mu_Z(x), \tag{2.6}$$

with $g \in L^2(\mathbb{R}^{2d}; \mu_Z; \mathbb{R}^{2d})$. As a consequence, the operator B, defined by

$$Bh := A^* A h = \int_{\mathbb{R}^{2d}} \nabla^T h(x) \nabla_1 K(x, \cdot) \, \mathrm{d}\mu_Z(x), \tag{2.7}$$

is a positive and trace class mapping from $\mathcal{H}_K$ to $\mathcal{H}_K$.

Operator representation

We define the operator A_N (empirical version of A) as

$$A_N h := \frac{1}{\sqrt{N}} \mathrm{Vec}(\{X_h(\mathbf{Z}^{(n)})\}_{n=1}^N), \quad h \in \mathcal{H}_K.$$

If the kernel $K \in C_b^3(\mathbb{R}^{2d} \times \mathbb{R}^{2d})$, the operator $A_N : \mathcal{H}_K \to \mathbb{R}^{2dN}$ is bounded linear. The adjoint operator A_N^* is

$$A_N^* W = \frac{1}{\sqrt{N}} W^T \mathbb{J} \nabla_1 K(\mathbf{Z}_N, \cdot),$$

with $W \in \mathbb{R}^{2dN}$, where $\mathbb{J} = \mathrm{diag}\{J, \cdots, J\}_{N \times N}$. And the operator B_N defined by

$$B_N h := A_N^* A_N h = \frac{1}{N} \nabla^T h(\mathbf{Z}_N) \nabla_1 K(\mathbf{Z}_N, \cdot), \tag{2.8}$$

is a positive and compact mapping $\mathcal{H}_K$ to $\mathcal{H}_K$.

Operator representation

For all $\lambda > 0$, the solutions of the optimization problems (5.1) and (2.3) exist and are unique:

$$\widehat{h}_{\lambda,N} := \operatorname*{arg\,min}_{h \in \mathcal{H}_K} \widehat{R}_{\lambda,N}(h) = \frac{1}{\sqrt{N}} (B_N + \lambda I)^{-1} A_N^* \mathbf{X}_{\sigma^2,N},$$

$$h_\lambda^* := (B + \lambda I)^{-1} A^* X_H.$$

The Differential Representer Theorem

Theorem

For every $\lambda > 0$, $\widehat{h}_{\lambda,N}$ can be represented as

$$\widehat{h}_{\lambda,N} = \sum_{i=1}^{N} \langle \widehat{\mathbf{c}}_i, \nabla_1 K(\mathbf{Z}^{(i)}, \cdot) \rangle,$$

with $\widehat{\mathbf{c}}_1, \ldots, \widehat{\mathbf{c}}_N \in \mathbb{R}^{2d}$ and $\langle \cdot, \cdot \rangle$ the Euclidean inner product in $\mathbb{R}^{2d}$. Moreover, if $\widehat{\mathbf{c}} \in \mathbb{R}^{2dN}$ is the vectorization of $(\widehat{\mathbf{c}}_1 | \cdots | \widehat{\mathbf{c}}_N)$, then

$$\widehat{\mathbf{c}} = (\nabla_{1,2} K(\mathbf{Z}_N, \mathbf{Z}_N) + \lambda N I)^{-1} \mathbb{J}^\top \mathbf{X}_{\sigma^2, N}.$$

The matrix $\nabla_{1,2} K(\mathbf{Z}_N, \mathbf{Z}_N)$ is the differential Gram matrix which is positive semidefinite.

How can the solution be unique?

Define the kernel of A:

$$\mathcal{H}_{\text{null}} := \{h \in \mathcal{H}_K \mid Ah = X_h = 0\} = \{f \in \mathcal{H}_K \mid \nabla h = 0\}.$$

It can be shown that $\widehat{h}_{\lambda,N} \in \mathcal{H}_{\text{null}}^{\perp}$. The uniqueness of the optimizer is due to the use of the regularization term:

- Let $\widehat{h}_{\lambda,N}$ and let $h \in \mathcal{H}_{\text{null}}$.
- $\widehat{h}_{\lambda,N}$ and $\widehat{h}_{\lambda,N} + h$ have the same Hamiltonian vector field associated but $\widehat{h}_{\lambda,N} + h$ is an empirical risk minimizer if and only if $h \equiv 0$.
- This is because

$$\widehat{R}_{\lambda,N}(\widehat{h}_{\lambda,N} + h) = \frac{1}{N}\sum_{n=1}^{N} \|X_{\widehat{h}_{\lambda,N}}(\mathbf{Z}^{(n)}) - \mathbf{X}_{\sigma^2}^{(n)}\|^2 + \lambda \left(\|\widehat{h}_{\lambda,N}\|_{\mathcal{H}_K}^2 + \|h\|_{\mathcal{H}_K}^2 \right)$$

Connection with Gaussian Posterior Mean Estimator

Step 1: Model the Hamiltonian H as a GP prior $\mathcal{GP}(0, K^{\theta})$.

Step 2: Maximize the log marginal likelihood
$-\log p(\mathbf{X}_{\sigma^2,N}|\mathbf{z}_N, \mathbf{x}_{\sigma^2,N}, \theta, \sigma^2)$.

Step 3: Make the prediction: For each $\mathbf{z}^* \in \mathbb{R}^{2d}$, $H(\mathbf{z}^*)$ satisfies

$$H(\mathbf{z}^*)|\mathbf{z}_N, \mathbf{x}_{\sigma^2,N} \sim \mathcal{N}(\overline{\phi}_N(\mathbf{z}^*), \overline{\Sigma}_N(\mathbf{z}^*)),$$

where

$$\overline{\phi}_N(\mathbf{z}^*) = K^{\widehat{\theta}}_{H,X_H}(\mathbf{z}^*, \mathbf{z}_N)(K^{\widehat{\theta}}_{X_H}(\mathbf{z}_N, \mathbf{z}_N) + \widehat{\sigma}^2 I_{2dN})^{-1}\mathbf{x}_{\sigma^2,N},$$
$$\overline{\Sigma}_N(\mathbf{z}^*) = K^{\widehat{\theta}}(\mathbf{z}^*, \mathbf{z}^*) - K^{\widehat{\theta}}_{H,X_H}(\mathbf{z}^*, \mathbf{z}_N)(K^{\widehat{\theta}}_{X_H}(\mathbf{z}_N, \mathbf{z}_N) + \widehat{\sigma}^2 I_{2dN})^{-1} K^{\widehat{\theta}}_{X_H,H}(\mathbf{z}_N, \mathbf{z}^*).$$

Connection:

$$\overline{\phi}_N = \widehat{h}_{\lambda,N} \iff \lambda = \frac{\sigma^2}{N}.$$

Online regression with kernels

The structure-preserving kernel estimator is

$$\widehat{h}_{\lambda,N} = \widehat{\mathbf{c}}_N \cdot \nabla_1 K(\mathbf{Z}_N, \cdot), \quad \text{with}$$

$$\widehat{\mathbf{c}}_N = (\nabla_{1,2} K(\mathbf{Z}_N, \mathbf{Z}_N) + \lambda N I)^{-1} \mathbb{J}^\top \mathbf{X}_{\sigma^2,N} =: \mathbf{K}_N^{-1} \mathbb{J}^\top \mathbf{X}_{\sigma^2,N}.$$

We now observe one more data point $(\mathbf{Z}, \mathbf{X})$. If $\lambda(N)N = C$,

$$\mathbf{K}_{N+1}^{-1} = \begin{bmatrix} \mathbf{K}_N^{-1} + \mathbf{K}_N^{-1}\mathbf{b}_N\mathbf{D}_N^{-1}\mathbf{b}_N^\top\mathbf{K}_N^{-1} & -\mathbf{K}_N^{-1}\mathbf{b}_N\mathbf{D}_N^{-1} \\ -\mathbf{D}_N^{-1}\mathbf{b}_N^\top\mathbf{K}_N^{-1} & \mathbf{D}_N^{-1} \end{bmatrix},$$

where $\mathbf{D}_N = \mathbf{A} - \mathbf{b}_N^\top \mathbf{K}_N^{-1} \mathbf{b}_N$ and the matrix $\mathbf{A} = \nabla_{1,2} K(\mathbf{Z}, \mathbf{Z}) + CI$.

- Deal with large training datasets in a cheap way.
- Easy to update the kernel estimator when new data comes in.

Error analysis

Convergence analysis

Estimation and approximation errors

$$\underbrace{\widehat{h}_{\lambda,N} - H}_{\text{Reconstruction error}} = \underbrace{\widehat{h}_{\lambda,N} - h^*_\lambda}_{\text{Estimation error}} + \underbrace{h^*_\lambda - H}_{\text{Approximation error}}$$

Approximation error: source condition. We assume that

$$H \in \Omega^\gamma_S := \{h \in \mathcal{H}_K \mid h = B^\gamma \psi, \psi \in \mathcal{H}_K, \|\psi\|_{\mathcal{H}_K} < S\}.$$

This is the source condition [FKRT23]. As the parameter γ increases, the functions in Ω^γ_S are smoother. The source condition implies that the approximation error can be bound using the RKHS norm as

$$\|h^*_\lambda - H\|_{\mathcal{H}_K} \leq \lambda^\gamma \|B^{-\gamma} H\|_{\mathcal{H}_K}.$$

Estimation error: Γ-convergence and probabilistic inequalities, Hanson–Wright inequality.

PAC bounds with fixed Tikhonov parameter

Theorem (PAC bounds of the total reconstruction error)

Suppose that $K \in C_b^3(\mathbb{R}^{2d} \times \mathbb{R}^{2d})$ *and* $H \in \Omega_S^\gamma$. *Then for every* $\varepsilon, \delta > 0$, *there exist* $\lambda > 0$ *and* $n \in \mathbb{N}_+$ *such that for all* $N > n$, *it holds that*

$$\mathbb{P}\left(\left\| \widehat{h}_{\lambda,N} - H \right\|_{\mathcal{H}_K} > \varepsilon \right) < \delta.$$

Convergence rates with adaptive Tikhonov parameter

Consider a dynamical Tikhonov parameter

$$\lambda \propto N^{-\alpha}, \quad \alpha > 0 \tag{3.1}$$

Theorem (Convergence rate of the total reconstruction error)

Suppose that $K \in C_b^3(\mathbb{R}^{2d} \times \mathbb{R}^{2d})$ *and* $H \in \Omega_S^\gamma$. *Then for all* $\alpha \in (0, \frac{1}{3})$, *and for any* $0 < \delta < 1$, *with probability as least* $1 - \delta$, *it holds that*

$$\|\widehat{h}_{\lambda,N} - H\|_{\mathcal{H}_K} \leq C(\gamma, \delta, \kappa)\ N^{-\min\{\alpha\gamma, \frac{1}{2}(1-3\alpha)\}},$$

where $C(\gamma, \delta, \kappa) = \max\left\{\|B^{-\gamma} H\|_{\mathcal{H}_K}, \sqrt{8\log(8/\delta)}d\kappa^3 \|H\|_{\mathcal{H}_K}\right\}$.

Convergence rates with coercivity condition

Coercivity condition: [FKRT23] There exists a constant $c_{\mathcal{H}_K} > 0$ such that

$$\|Ah\|^2_{L^2(\mu_Z)} = \|X_h\|^2_{L^2(\mu_Z)} \geq c_{\mathcal{H}_K} \|h\|^2_{\mathcal{H}_K}, \quad \forall\, h \in \mathcal{H}_K. \tag{3.2}$$

Theorem (Convergence rate of the total reconstruction error)

Suppose that $K \in C^3_b(\mathbb{R}^{2d} \times \mathbb{R}^{2d})$ and $H \in \Omega^\gamma_S$. Under coercivity condition (3.2), for all $\alpha \in (0, \frac{1}{2})$, and for any $0 < \delta < 1$, with probability as least $1 - \delta$, it holds that

$$\|\widehat{h}_{\lambda,N} - H\|_{\mathcal{H}_K} \leq C(\gamma, \delta, \sigma, \kappa, c_{HW}, c_{\mathcal{H}_K})\ N^{-\min\{\alpha\gamma, \frac{1}{2}(1-2\alpha)\}},$$

where

$$C(\gamma, \delta, \sigma, \kappa, c_{HW}, c_{\mathcal{H}_K}) = \max\left\{ \|B^{-\gamma}H\|_{\mathcal{H}_K}, \frac{\sigma\kappa}{\sqrt{2d}}\left(1 + \sqrt{\frac{1}{c_{HW}}\log(4/\delta)}\right), \sqrt{8\log(8/\delta)}d\kappa^2\left(2 + \frac{\kappa}{\sqrt{c_{\mathcal{H}_K}}}\right)\|H\|_{\mathcal{H}_K} \right\}.$$

Numerical experiments

Gaussian kernel:

$$K_\eta(x, y) = \exp\left(-\frac{\|x-y\|^2}{\eta^2}\right).$$

Dynamical Tikhonov regularization parameter:

$$\lambda = cN^{-\alpha}.$$

Estimator:

$$\widehat{h}_{\lambda,N} = \nabla_1 K^\top(\mathbf{z}_N, \cdot)(\nabla_{1,2} K(\mathbf{z}_N, \mathbf{z}_N) + \lambda N I)^{-1} \mathbb{J}^T \mathbf{x}_{\sigma^2,N}.$$

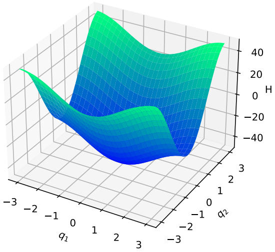
In the numerical experiments, we shall fix $\alpha = 0.4$ and search the parameters η and c.

Double pendulum

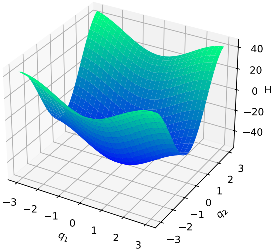
Consider the following Hamiltonian function

$$
\begin{aligned}
H(\theta_1, \theta_2, p_1, p_2) = \;& p_1\dot{\theta}_1 + p_2\dot{\theta}_2 + \frac{1}{2} mgl(3\cos\theta_1 + \cos\theta_2) \\
& - \frac{1}{6} ml^2(\dot{\theta}_2^{\,2} + 4\dot{\theta}_1^{\,2} + 3\dot{\theta}_1\dot{\theta}_2 \cos(\theta_1 - \theta_2)).
\end{aligned}
$$

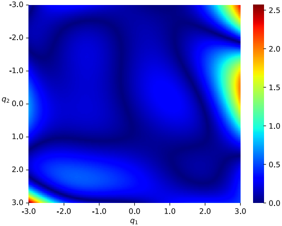
Double pendulum



(a)



(b)



(c)

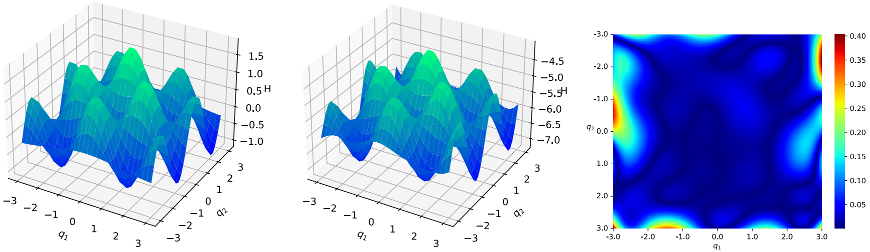
Figure: Double pendulum ($p_1 = p_2 = 0, N = 200$): (a) Groundtruth Hamiltonian (b) Learned Hamiltonian (c) Mismatch error after vertical shift.

Highly non-convex potential well

Consider the Hamiltonian function

$$
\begin{aligned}
H(q_1, q_2, p_1, p_2) = {} & \frac{1}{2}(p_1^2 + p_2^2) \\
& + \sin\left(\frac{2\pi}{3} \cdot q_1\right) \cos\left(\frac{2\pi}{3} \cdot q_2\right) + \frac{\sin(\sqrt{q_1^2 + q_2^2})}{\sqrt{q_1^2 + q_2^2}}.
\end{aligned}
$$

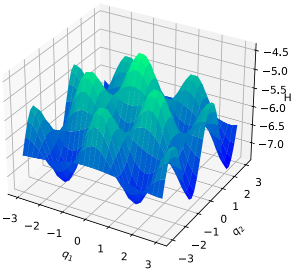
Highly non-convex potential well



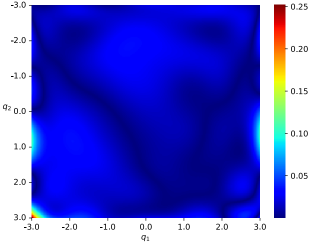
(a) (b) (c)

Figure: Highly non-convex potential well ($p_1 = p_2 = 0, N = 500$): (a) Groundtruth Hamiltonian (b) Learned Hamiltonian (c) Mismatch error after vertical shift.

Highly non-convex potential well



(a)



(b)

Figure: Highly non-convex potential well ($p_1 = p_2 = 0, N = 1500$): (a) Learned Hamiltonian (b) Mismatch error after vertical shift.

Learning Hamiltonian systems on manifolds

Motivation: The phase spaces of Hamiltonian systems are, in general:

- Symplectic manifolds (e.g. cotangent bundles)
- Poisson manifolds (e.g. Lie-Poisson)

that we shall endow with a Riemannian metric.

Observation data regime The random samples consist of

$$\{\mathbf{Z}_N, \mathbf{X}_{\sigma^2,N}\} := \{(\mathbf{Z}^{(n)})_{n=1}^N, (\mathbf{X}_{\sigma^2}^{(n)})_{n=1}^N\} \xrightarrow{realization} \{\mathbf{z}_N, \mathbf{x}_{\sigma^2,N}\}.$$

- The noisy vector fields $\mathbf{X}_{\sigma^2}^{(n)} = X_H(\mathbf{Z}^{(n)}) + \boldsymbol{\varepsilon}^{(n)}$ where $\mathbf{Z}^{(n)}$ are IID random variables on a symplectic manifold M with distribution μ_Z and $\boldsymbol{\varepsilon}^{(n)}$ are IID random variables on $T_{Z^{(n)}}M$ with $\mathbb{E}[\boldsymbol{\varepsilon}^{(n)}] = \mathbf{0}$ and $\mathbb{E}[\boldsymbol{\varepsilon}^{(n)}]^2 = \sigma^2 I_{2d}$.

Learning problem on manifolds

Optimization problem: We consider the following optimization using the regularized empirical risk

$$\widehat{h}_{\lambda,N} := \operatorname*{arg\,min}_{h\in\mathcal{H}_K} \widehat{R}_{\lambda,N}(h), \tag{5.1}$$

$$\widehat{R}_{\lambda,N}(h) := \frac{1}{N}\sum_{n=1}^{N} \|X_h(\mathbf{Z}^{(n)}) - \mathbf{X}_{\sigma^2}^{(n)}\|_g^2 + \lambda\|h\|_{\mathcal{H}_K}^2. \tag{5.2}$$

The corresponding optimization problem for the regularized statistical risk is

$$h_\lambda^* := \operatorname*{arg\,min}_{h\in\mathcal{H}_K} R_\lambda(h), \tag{5.3}$$

$$R_\lambda(h) := \|X_h - X_H\|_{L^2(\mu_Z)}^2 + \lambda\|h\|_{\mathcal{H}_K}^2 + \sigma^2. \tag{5.4}$$

High-order differentials and the space $C_b^s(M)$

If $f : M \to \mathbb{R}$ is in C^k class, we define the k-order differential of f denoted as $D^k f : T^k M \to \mathbb{R}$ inductively to be the differential of $D^{k-1} f : T^{k-1} M \to \mathbb{R}$.

$C_b^s(M)$ is the set of functions in C^s class with bounded s-order differentials

$$C_b^s(M) := \{ f \in C^s(M) \mid \|f\|_\infty + \sum_{k=1}^{s} \|D^k f\|_\infty < \infty \},$$

where $\|f\|_\infty := \sup_{x \in M} |f(x)|$ and

$$\|D^k f\|_\infty := \sup_{y \in T^{k-1} M} \ \sup_{v \in T_y T^{k-1} M} \frac{\left| D^k f(y) \cdot v \right|}{\|v\|_{k-1}}.$$

$\|v\|_{k-1}$ stands for the norm of v in the tangent space $T_y T^{k-1} M$. If M is a Riemannian manifold with metric g, this norm can be induced by g.

Learning on symplectic manifolds

Compatible structure: We equip the manifold M with both a symplectic form ω and a Riemannian metric g. Then, we can define a map $J : TM \to TM$ given by

$$Jv := \omega_x^\sharp \left(g_x^\flat(v) \right), \quad \forall\, x \in M, v \in T_xM, \tag{5.5}$$

where $\omega^\sharp : T^*M \to TM$ and $g^\flat : TM \to T^*M$ are the bundle isomorphisms determined by the symplectic form ω and the Riemannian metric g, respectively.

Hamiltonian vector fields: $X_h = \omega^\sharp(dh) = \omega^\sharp \left(g^\flat(\nabla h) \right) = J\nabla h.$

Warning: we are not imposing Kähler despite the notation.

Learning on Poisson manifolds

The Poisson tensor: $(P, \{\cdot,\cdot\})$ be a Poisson manifold. The Poisson tensor is the contravariant anti-symmetric two-tensor $B : T^*P \times T^*P \to \mathbb{R}$, defined by

$$B(z)\,(\alpha_z, \beta_z) = \{F, G\}(z), \text{ where } \mathbf{d}F(z) = \alpha_z \text{ and } \mathbf{d}G(z) = \beta_z \in T_z^*P.$$

Compatible structure: $B^\sharp : T^*P \to TP$ vector bundle map associated to the B by $B(z)\,(\alpha_z, \beta_z) = \alpha_z \cdot B^\sharp(z)(\beta_z)$. Define the vector bundle map $J : TP \to TP$ by

$$J(z)v := B^\sharp(z)\left(g^\flat(z)(v)\right), \quad \forall\, z \in P, v \in T_zP, \tag{5.6}$$

Hamiltonian vector fields: $X_h = B^\sharp(\mathbf{d}h) = B^\sharp(g^\flat(\nabla h)) = J\nabla h.$

Poisson degeneracy

Important difference between symplectic and Poisson manifolds is that the Poisson tensor can of varying and non-constant rank. This is always the case when the Poisson algebra has a center

$$\mathcal{C}(P) = \{C \in C^{\infty}(P) \mid \{C, F\} = 0, \text{ for all } F \in C^{\infty}(P)\},$$

Elements in $\mathcal{C}(P)$ are called Casimirs. If $C \in C^{\infty}(P)$ then C is constant along the flow of all Hamiltonian vector fields, equivalently, $X_C = 0$.

Hamiltonians are defined only up to Casimirs.

Example I: The rigid body

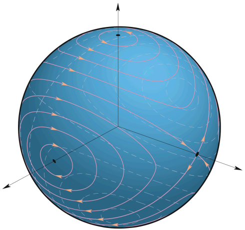
The rigid body satisfies a Lie-Poisson equation on $\mathfrak{so}(3)^* \simeq \mathbb{R}^3$ determined by the Hamiltonian function

$$H(\Pi) = \frac{1}{2}\Pi^\top \mathbb{I}^{-1}\Pi,$$

where $\mathbb{I} = \mathrm{diag}\{I_1, I_2, I_3\}$ is a diagonal matrix. The Poisson bracket is

$$\{F, K\}(\Pi) = -\Pi \cdot (\nabla F \times \nabla K).$$

$C(\Pi) = \|\Pi\|^2$ is a Casimir function of the Poisson algebra. $\Phi : \mathbb{R} \to \mathbb{R}$ differentiable function implies the function $\Phi \circ C$ is also a Casimir.



Taken from [MR99]

Example II: The underwater vehicle [Leo97]

The underwater vehicle has Lie-Poisson dynamics on $\mathfrak{so}(3)^* \times \mathbb{R}^{3^*} \times \mathbb{R}^{3^*}$ determined by the Hamiltonian function

$$H(\Pi, Q, \Gamma) = \frac{1}{2}\left(\Pi^\top A \Pi + 2\Pi^\top B^\top Q + Q^\top C Q - 2mg\left(\Gamma \cdot r_G\right)\right),$$

The Poisson bracket on $\mathfrak{so}(3)^* \times \mathbb{R}^{3^*} \times \mathbb{R}^{3^*}$ is

$$\{F, K\}(\Pi, Q, \Gamma) = \nabla F^\top \Lambda(\Pi, Q, \Gamma) \nabla K,$$

where the Poisson tensor Λ is given by

$$\Lambda(\Pi, Q, \Gamma) = \begin{pmatrix} \hat{\Pi} & \hat{Q} & \hat{\Gamma} \\ \hat{Q} & 0 & 0 \\ \hat{\Gamma} & 0 & 0 \end{pmatrix}.$$

Casimir functions:

$$C_1(\Pi, Q, \Gamma) = Q \cdot \Gamma, \quad C_2(\Pi, Q, \Gamma) = \|Q\|^2, \quad C_3(\Pi, Q, \Gamma) = \|\Gamma\|^2,$$
$$C_4(\Pi, Q, \Gamma) = \Pi \cdot Q, \quad C_5(\Pi, Q, \Gamma) = \Pi \cdot \Gamma, \quad C_6(\Pi, Q, \Gamma) = \|\Pi\|^2.$$

Well-posedness of the optimization problems

Boundedness condition: The compatible structure J satisfies

$$g(Jv, Jv) \leq \gamma(x) g(v, v), \quad \text{for all } v \in T_x M, \tag{5.7}$$

where γ is a positive bounded function on P. The boundedness of J gives us the boundedness of the operators A and A_N.

Boundedness of A: Define the operator:

$$Ah := X_h = J\nabla h, \quad h \in \mathcal{H}_K,$$

If kernel $K \in C_b^3(M \times M)$, then the boundedness condition implies that $A : \mathcal{H}_K \longrightarrow L^2(M, \mu_Z)$ is bounded linear. The operator $Q : \mathcal{H}_K \longrightarrow \mathcal{H}_K$, defined by

$$Qh := A^* Ah = \int_M g(X_{K_\cdot}(x), X_h(x)) \, d\mu_Z(x). \tag{5.8}$$

is positive trace class.

Operator representations of the minimizers

Let $A_N : \mathcal{H}_K \to T_{Z_N}P := \Pi_{i=1}^N T_{Z^{(i)}}P$ as

$$A_N h := \frac{1}{\sqrt{N}} X_h(\mathbf{Z}_N) := \frac{1}{\sqrt{N}} \mathrm{Vec}\left(J(\mathbf{Z}^{(1)}) \nabla h(\mathbf{Z}^{(1)}) | \cdots | J(\mathbf{Z}^{(N)}) \nabla h(\mathbf{Z}^{(N)}) \right).$$

Proposition

If boundedness assumption holds then $A_N : \mathcal{H}_K \to T_{Z_N}P$ is bounded. The adjoint operator $A_N^* : T_{Z_N}P \to \mathcal{H}_K$ of A_N is finite rank and given by

$$A_N^* W = \frac{1}{\sqrt{N}} g_N(W, X_{K_\cdot}(\mathbf{Z}_N)), \quad W \in T_{Z_N}P.$$

The operator Q_N defined by

$$Q_N h := A_N^* A_N h = \frac{1}{N} g_N(X_h(\mathbf{Z}_N), X_{K_\cdot}(\mathbf{Z}_N)), \quad h \in \mathcal{H}_K, \tag{5.9}$$

is a positive-semidefinite compact operator.

Operator representations:

$$h_\lambda^* := (Q + \lambda I)^{-1} A^* X_H.$$
$$\widehat{h}_{\lambda,N} := \frac{1}{\sqrt{N}} (Q_N + \lambda I)^{-1} A_N^* \mathbf{X}_{\sigma^2,N}.$$

Kernel representations of the minimizers: define the generalized differential Gram matrix $G_N : T_{Z_N}P \to T_{Z_N}P$ as

$$G_N \mathbf{c} := X_{g_N(\mathrm{c}, X_{K.(\mathrm{Z}_N)})}(\mathbf{Z}_N), \quad \mathbf{c} \in T_{Z_N}P.$$

In even dimensional Euclidean spaces reduces to the usual differential Gram matrix $\mathbb{J}_{\mathrm{can}} \nabla_{1,2} K(\mathbf{Z}_N, \mathbf{Z}_N) \mathbb{J}_{\mathrm{can}}^\top$.

Property: Given a Mercer kernel $K \in C_b^3(M \times M)$, the general differential Gram matrix $G_N : T_{Z_N}M \to T_{Z_N}M$ is symmetric and positive semidefinite.

Differential Representer Theorem on Poisson manifolds

Theorem

Suppose $K \in C_b^3(P \times P)$ *and* J *is bounded. Then, can be represented as*

$$\widehat{h}_{\lambda,N} = g_N(\widehat{\mathbf{c}}, X_{K.}(\mathbf{Z}_N)),$$

where $\widehat{\mathbf{c}} \in T_{Z_N}P$ *is given by*

$$\widehat{\mathbf{c}} = (G_N + \lambda N I)^{-1}\mathbf{X}_{\sigma^2,N}.$$

What about Poisson degeneracy?

Define the kernel of A as:

$$\mathcal{H}_{\text{null}} := \{h \in \mathcal{H}_K \mid Ah = J\nabla h = 0\}.$$

$\mathcal{H}_{\text{null}}$ is a closed subspace of $\mathcal{H}_K$ and hence $\mathcal{H}_K$ can be decomposed as

$$\mathcal{H}_K = \mathcal{H}_{\text{null}} \oplus \mathcal{H}_{\text{null}}^{\perp},$$

This decomposition and the expression of the kernel estimator implies

$$\widehat{h}_{\lambda,N} \in \mathcal{H}_{\text{null}}^{\perp}.$$

Why is the estimator $\widehat{h}_{\lambda,N}$ unique and not up to Casimir functions? The answer is in the use of the regularization term. Let $h \in \mathcal{H}_{\text{null}}$ then $\widehat{h}_{\lambda,N}$ and $\widehat{h}_{\lambda,N} + h$ have the same Hamiltonian vector field associated, but it is easy to show that $\widehat{h}_{\lambda,N} + h$ is a minimizer if and only if $h \equiv 0$. This is because:

$$\widehat{R}_{\lambda,N}(\widehat{h}_{\lambda,N} + h) = \frac{1}{N}\sum_{n=1}^{N} \|X_{\widehat{h}_{\lambda,N}}(\mathbf{Z}^{(n)}) - \mathbf{X}_{\sigma^2}^{(n)}\|^2 + \lambda\left(\|\widehat{h}_{\lambda,N}\|_{\mathcal{H}_K}^2 + \|h\|_{\mathcal{H}_K}^2\right).$$

What else?

- Availability of coordinate expressions
- Very similar error bounds and convergence rates.

Example: $\widehat{h}_{\lambda,N}$ in the Lie-Poisson case

Equip the dual Lie algebra $\mathfrak{g}^*$ with the Lie-Poisson bracket $\{\cdot,\cdot\}_+$:

$$\{F,G\}_+(\mu) = \left\langle \mu, \left[\frac{\delta F}{\delta \mu}, \frac{\delta G}{\delta \mu}\right]\right\rangle.$$

The Lie-Poisson system associated with a Hamiltonian $H : \mathfrak{g}^* \to \mathbb{R}$ is

$$\dot{\mu} = X_H(\mu) = \mathrm{ad}^*_{\frac{\delta H}{\delta \mu}} \mu.$$

The generalized differential Gram matrix G_N is

$$G_N \mathbf{c} = X_{g_N(\mathrm{c}, X_{K_\cdot}(\mathrm{Z}_N))}(\mathbf{Z}_N) = X_{\mathrm{c}^\top \mathbb{J}_N \nabla_1 K(\mathrm{Z}_N,\cdot)}(\mathbf{Z}_N) = \mathbb{J}_N \nabla_{1,2} K(\mathbf{Z}_N, \mathbf{Z}_N)\mathbb{J}_N^\top \mathbf{c},$$

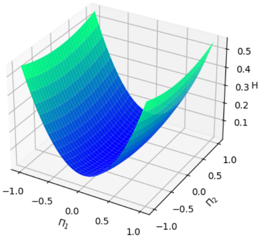
for all $\mathbf{c} \in T_{Z_N}\mathfrak{g}^*$. Therefore,

$$\widehat{h}_{\lambda,N} = \mathbf{X}^\top_{\sigma^2,N}\left(\mathbb{J}_N \nabla_{1,2} K(\mathbf{Z}_N, \mathbf{Z}_N)\mathbb{J}_N^\top + \lambda N I\right)^{-1} \mathbb{J}_N \nabla_1 K(\mathbf{Z}_N, \cdot),$$

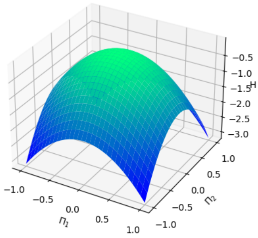
where the compatible structure J defined is given by

$$J(\mu)\xi = \mathrm{ad}^*_{\langle \xi, \cdot\rangle} \mu, \qquad \text{for all } \xi \in \mathfrak{g}^*.$$

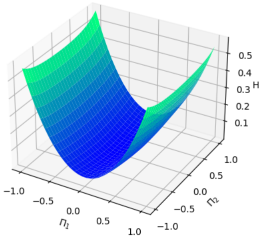
Numerical illustration: Rigid body



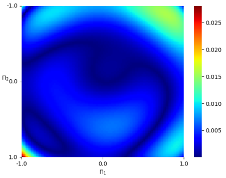
(a)



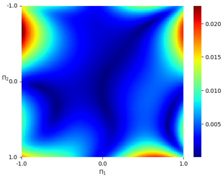
(b)



(c)

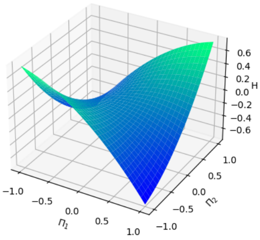


(d)

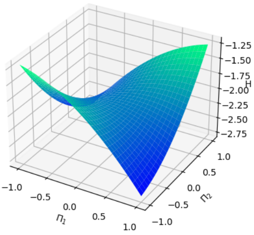


(e)

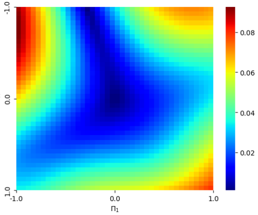
Numerical illustration: Underwater vehicle



(a)



(b)



(c)

Figure: Underwater Vehicle: (a) Groundtruth Hamiltonian (b) Learned Hamiltonian with $N = 400$ (c) Error of the predicted Hamiltonian vector field

Perspectives

- Argumentwise invariant kernels and momentum map preservation.
- What about time series data?
- Use universality arguments and develop universality for kernels on manifolds.

References I

Christopher Eldred, François Gay-Balmaz, Sofiia Huraka, and Vakhtang Putkaradze.
LiePoisson Neural Networks (LPNets): Data-based computing of Hamiltonian systems with symmetries.
Neural Networks, 173:106162, 2024.

Jinchao Feng, Charles Kulick, Yunxiang Ren, and Sui Tang.
Learning particle swarming models from data with {G}aussian processes.
Mathematics of Computation, 2023.

Pengzhan Jin, Zhen Zhang, Ioannis G Kevrekidis, and George Em Karniadakis.
Learning Poisson systems and trajectories of autonomous systems via Poisson neural networks.
IEEE Transactions on Neural Networks and Learning Systems, 2022.

N. E. Leonard.
Stability of a bottom-heavy underwater vehicle.
Automatica, 33(3):331–346, 1997.

Jerrold E. Marsden and Tudor S. Ratiu.
Introduction to mechanics and symmetry.
Springer-Verlag, New York, second edition, 1999.