AEROSPACE ENGINEERING
UNIVERSITY OF MICHIGAN

Enforcing physical structure in Bayesian learning of dynamical systems: stability and energy conservation

Nick Galioto, Alex Gorodetsky – University of Michigan

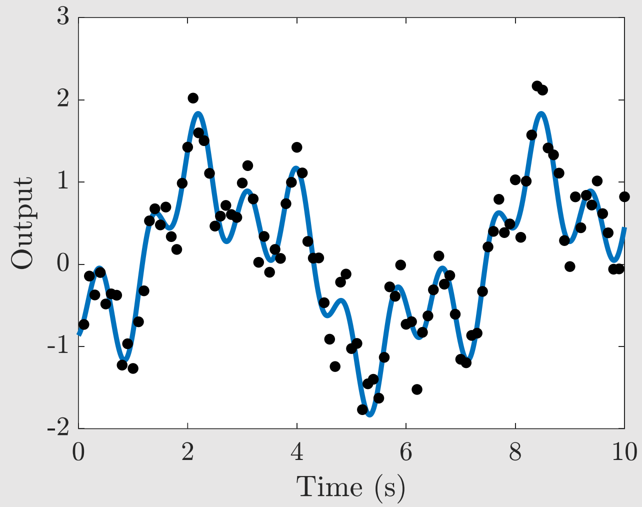
Harsh Sharma, Boris Kramer – University of California, San Diego

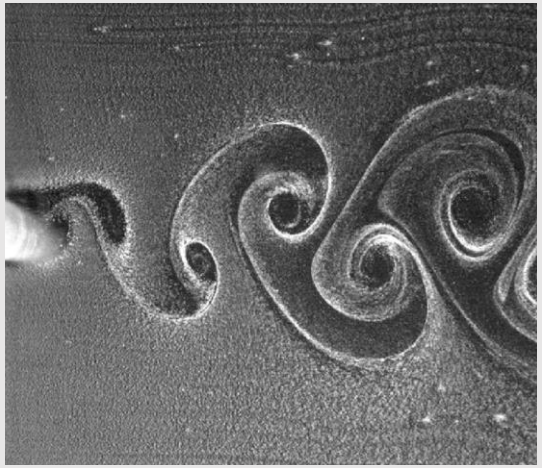
September 26, 2022

SIAM Conference on Mathematics of Data Science

Motivation

- **Goal:** learn a model of a dynamical system from time-series data
- Two primary design choices in system identification:
  - Model parameterization: neural networks, basis expansions, kernel expansions
  - **Objective function**: least squares, regularization, etc.
- A good model structure will:
  - **Enforce known physics**
  - Reduce data requirements and fill in for missing data
- A good objective will:
  - Be robust to sparse and noisy data
  - Handle model inadequacy
  - Generalize well beyond training data

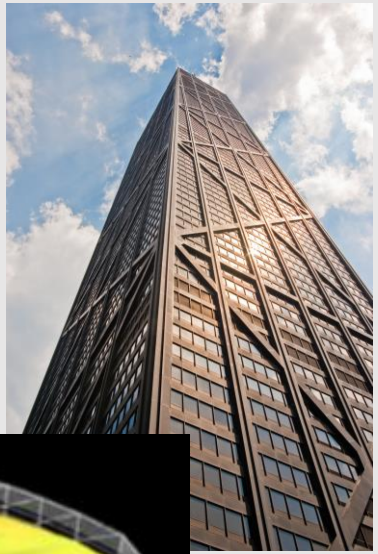


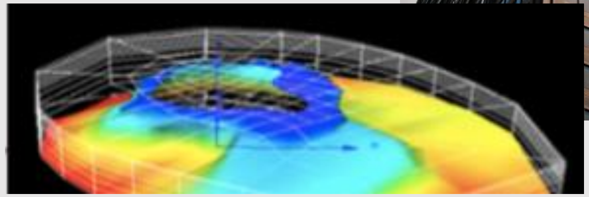


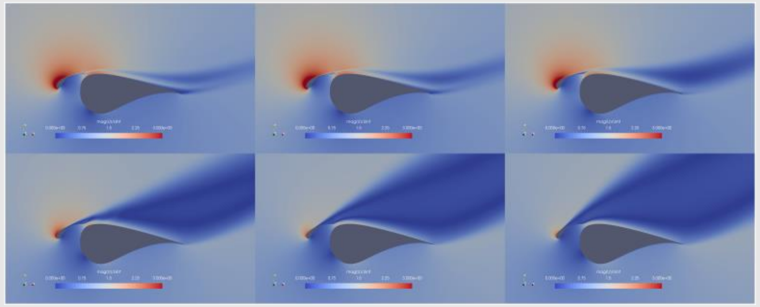
Imperfectly known models

Oftentimes, domain knowledge can produce reliable models, but problem-specific parameters may still be unknown

- Common in fields like structural dynamics and systems biology (material properties, kinetic parameters, etc.)
- Data can be expensive or challenging to collect
- Need to find accurate estimates and quantify uncertainty







Contributions

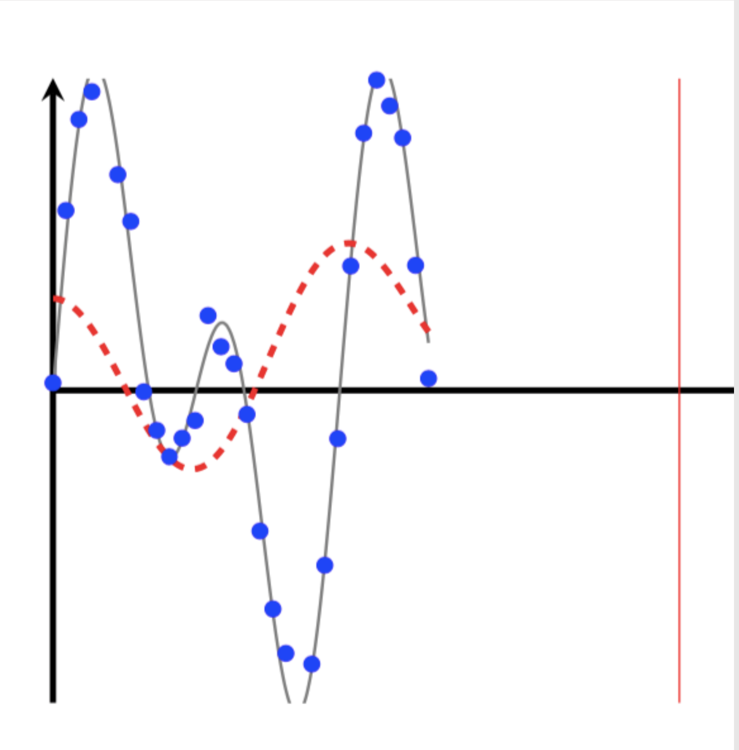
Present a system ID algorithm that can:
- Structurally embed physics constraints
- Handle measurement, model, and parameter uncertainty and their interaction
- Accurately identify parameters from sparse and noisy data
- Quantify model uncertainty

We seek probabilistic predictions to quantify uncertainty

Probabilistic Prediction

$$P(\text{ value is } x \mid \text{data, information })$$

- Data: time series, noisy and sparse
- Information: conservation of energy (Hamiltonian system)



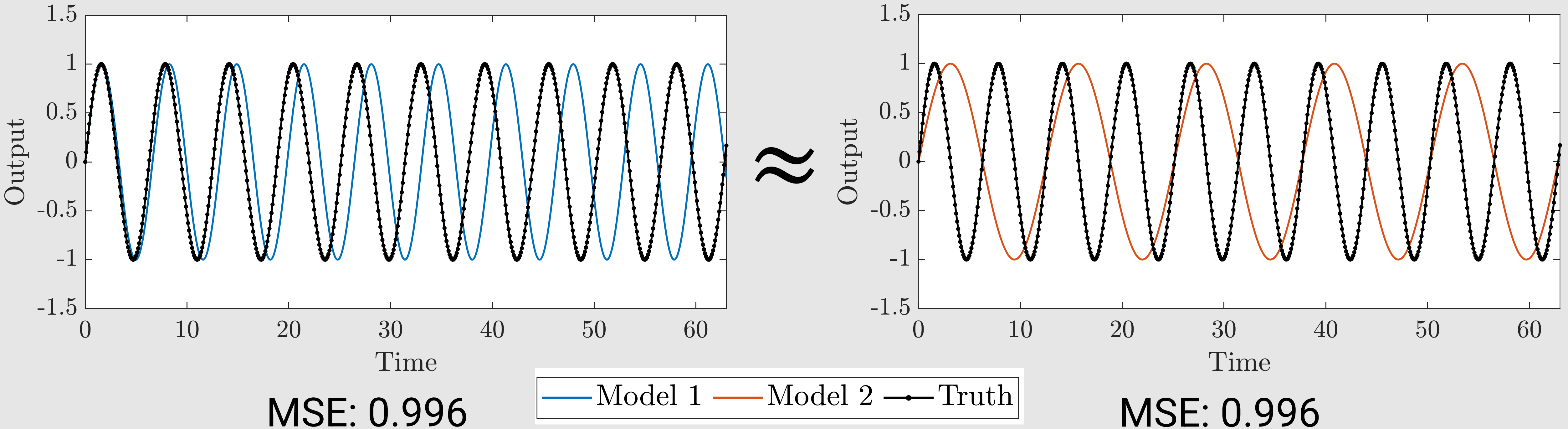
Outline

1. Existing approaches
2. Probabilistic formulation
3. Algorithm/Marginal likelihood
4. Hamiltonian Systems
5. Results
6. Takeaways

What’s wrong with the least squares objective?

One aspect: the least squares error metric can induce an undesirable ranking of dynamical models

The accumulation of small model errors is given equal weight as large model error



MSE: 0.996

MSE: 0.996

How can we design an objective that prioritizes Model 1 over Model 2?

Existing approaches

Least squares-based objective functions

(a) Assumes perfect model

$$J(\theta) = \frac{1}{n}\sum_{k=1}^{n} \|y_k - h(x(t_k), \theta)\|_2^2 \quad \text{s.t.} \quad \frac{\mathrm{d}x}{\mathrm{d}t} = f(t, x; \theta)$$

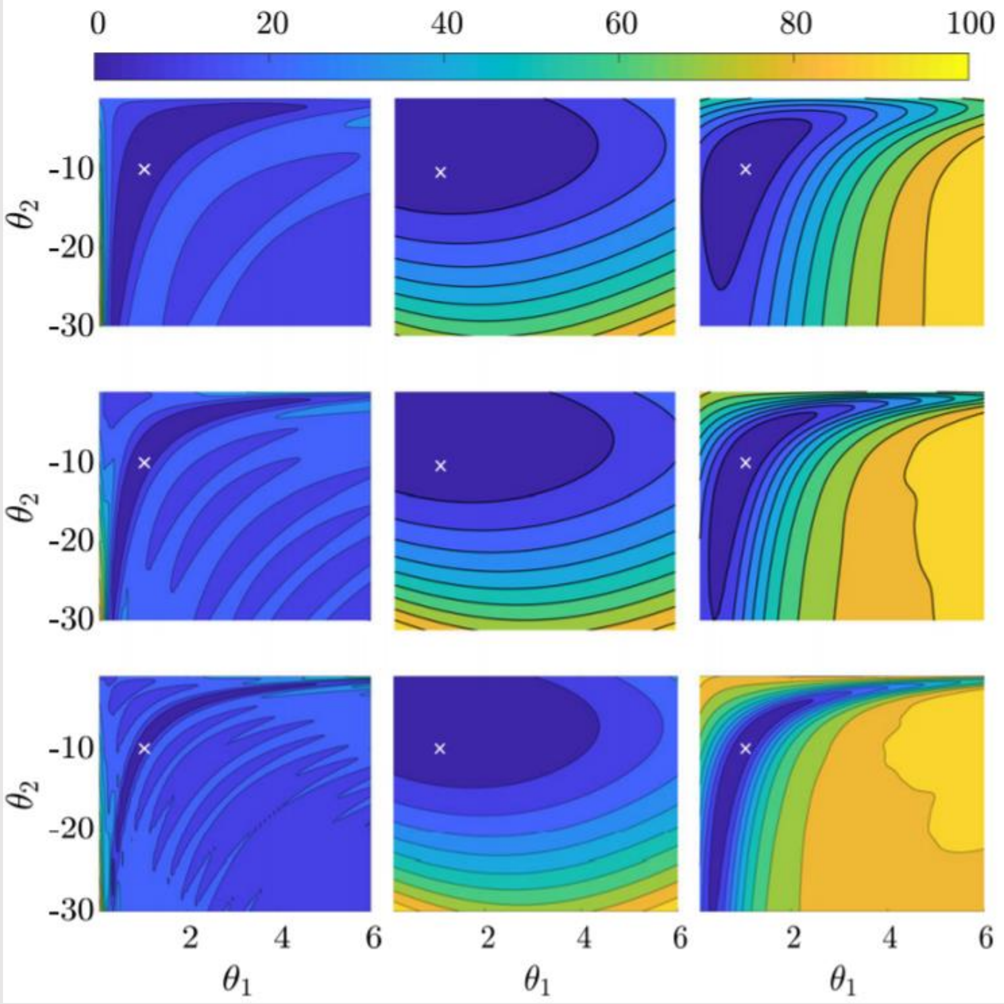
(b) Assumes noiseless measurements

$$J(\theta) = \frac{1}{n}\sum_{k=1}^{n} \|y_k - \Psi(y_{k-1}; \theta)\|_2^2$$

(c) Noisy measurements + model error (process noise)

- Optimal combination of (a) and (b)

measurements



(a) (b) (c)

	(a)	(b)	(c)
Steep optimization surfaces without plateaus	✓	✗	✓
Smooths local minima	✗	✓	✓
Increased confidence with data	✓	✗	✓

Outline

Probabilistic formulation: hidden Markov model

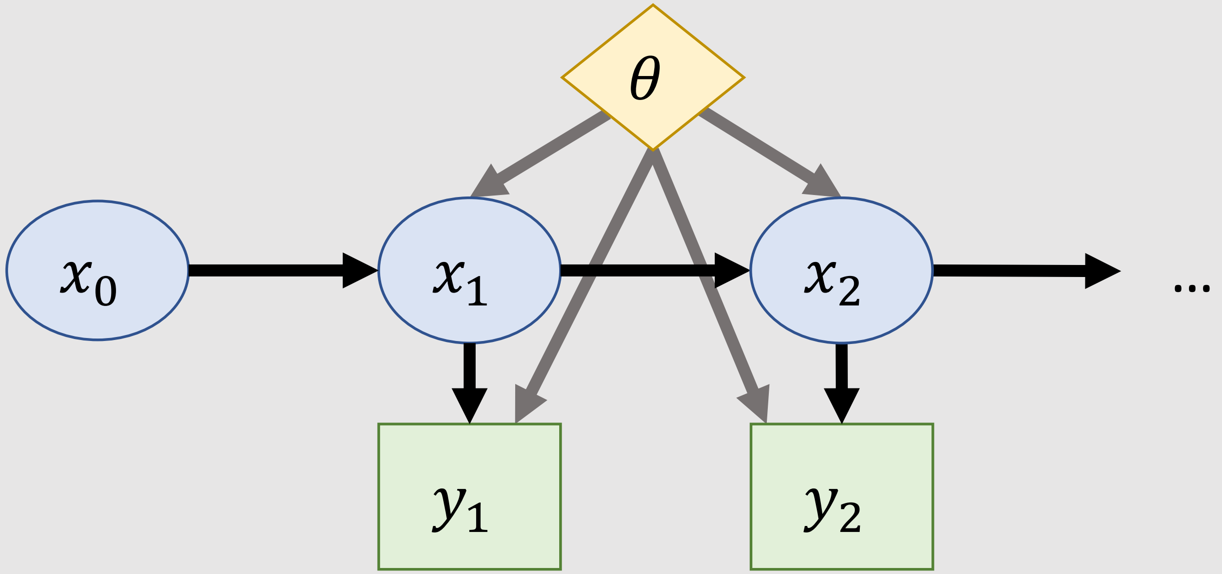
Joint parameter-state estimation with stochastic dynamics

$$X_k \in \mathbb{R}^{d_x}, \qquad Y_k \in \mathbb{R}^{d_y}, \qquad \theta = (\theta_\Psi, \theta_h, \theta_\Sigma, \theta_\Gamma) \in \mathbb{R}^{d_\theta}$$

$$X_k = \Psi(X_{k-1}, u_{k-1}, \theta_\Psi) + \xi_k; \quad \xi_k \sim \mathcal{N}\big(0, \Sigma(\theta_\Sigma)\big)$$

$$Y_k = h(X_k, \theta_h) + \eta_k; \qquad \eta_k \sim \mathcal{N}\big(0, \Gamma(\theta_\Gamma)\big)$$

The process noise term ξ_k accounts for model error
- Parameter error
- Integration error
- Insufficient model expressiveness



1. Parameter Uncertainty
2. Model Uncertainty
3. Measurement Uncertainty

Posterior flow chart

Log Joint Likelihood

$$\log \mathcal{L}(\theta;\, \mathcal{X}_n, \mathcal{Y}_n) \propto -\frac{1}{2}\sum_{k=1}^{n} \|y_k - h(x_k, \theta_h)\|^2_{\Gamma(\theta_\Gamma)} - \frac{1}{2}\sum_{k=1}^{n} \|x_k - \Psi(x_{k-1}, \theta_\Psi)\|^2_{\Sigma(\theta_\Sigma)}$$

Deterministic dynamics:
$x_k = \Psi(x_{k-1})$

$$\log \mathcal{L}(\theta;\, \mathcal{Y}_n) \propto -\frac{1}{2}\sum_{k=1}^{n} \left\|y_k - h\left(\Psi^k(x_0, \theta_\Psi), \theta_h\right)\right\|^2$$

- ODE-Net; Chen et al., 2018
- PDE-Net; Long et al., 2018
- UDE; Rackauckas et al., 2019

Identity observations:
$y_k = x_k$

$$\log \mathcal{L}(\theta;\, \mathcal{Y}_n) \propto -\frac{1}{2}\sum_{k=2}^{n} \|y_k - \Psi(y_{k-1}, \theta_\Psi)\|^2$$

- DMD; Schmid, 2010
- SINDy; Brunton et al., 2019
- Hamiltonian NN; Greydanus et al., 2019

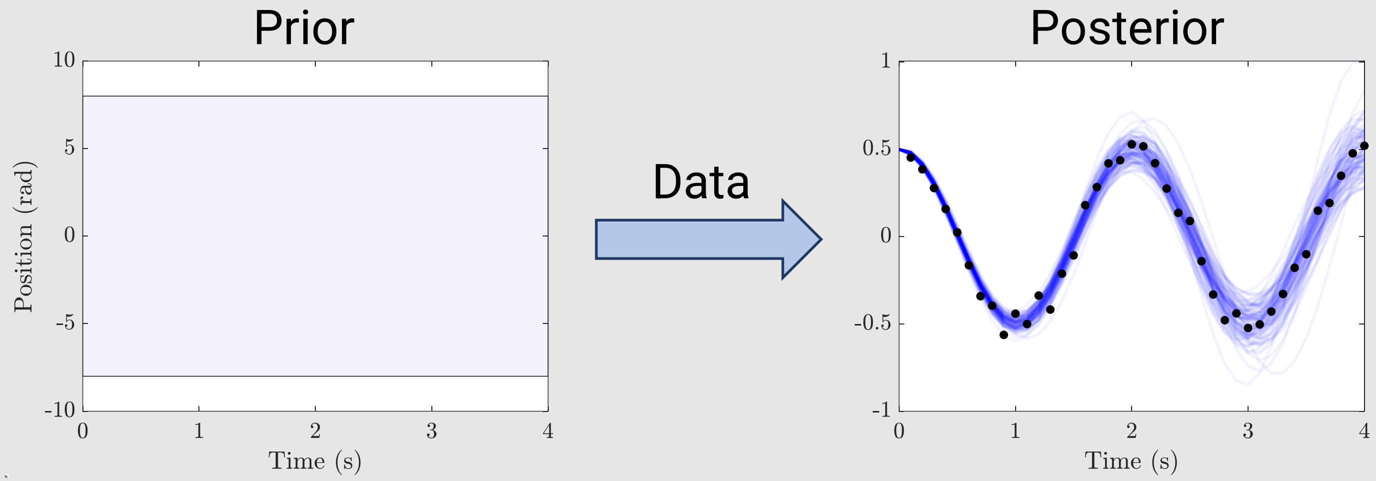
Chen, R. T., Rubanova, Y., Bettencourt, J., & Duvenaud, D. K. (2018). Neural ordinary differential equations. *Advances in neural information processing systems*, *31*.
Long, Z., Lu, Y., Ma, X., & Dong, B. (2018, July). Pde-net: Learning pdes from data. In *International Conference on Machine Learning* (pp. 3208-3216).
Rackauckas, C., Ma, Y., Martensen, J., Warner, C., Zubov, K., Supekar, R., ... & Edelman, A. (2020). Universal differential equations for scientific machine learning. *arXiv preprint arXiv:2001.04385*.
Schmid, P. J. (2010). Dynamic mode decomposition of numerical and experimental data. *Journal of fluid mechanics*, *656*, 5-28.
Brunton, S. L., Proctor, J. L., & Kutz, J. N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the national academy of sciences*, *113*(15), 3932-3937.
Greydanus, S., Dzamba, M., & Yosinski, J. (2019). Hamiltonian neural networks. *Advances in Neural Information Processing Systems*, *32*.

Outline

1. Existing approaches
2. Probabilistic formulation
3. **Algorithm/Marginal likelihood**
4. Hamiltonian Systems
5. Results
6. Takeaways

Bayesian Inference

- Goal: compute $p(\theta|\mathcal{Y}_n)$ where $\mathcal{Y}_n = (y_1, y_2, \ldots, y_n)$
- Bayes' rule: $p(\theta|\mathcal{Y}_n) = \frac{\mathcal{L}(\theta; \mathcal{Y}_n)p(\theta)}{p(\mathcal{Y}_n)}$



- Due to uncertainty in the states, we can only access the joint likelihood: $\mathcal{L}(\theta, \mathcal{X}_n;\ \mathcal{Y}_n)$
- To get the marginal likelihood, we must evaluate the integral

$$\mathcal{L}(\theta;\ \mathcal{Y}_n) = \int \mathcal{L}(\theta;\ \mathcal{X}_n, \mathcal{Y}_n)d\mathcal{X}_n$$

Marginal Markov Chain Monte Carlo (Särkkä, 2013)

Särkkä, S. (2013). *Bayesian filtering and smoothing* (No. 3). Cambridge university press.

1. **for** $i = 1, \ldots, N$
2. Propose sample θ
   Evaluate posterior: $p(\theta|\mathcal{Y}_n) = p(\theta) \prod_{k=1}^{n} \mathcal{L}_k(\theta; \mathcal{Y}_k)$
3. **for** $k = 0, \ldots, n-1$
4. Predict: $p(X_{k+1}|\mathcal{Y}_k, \theta) = \int p(X_{k+1}|X_k, \theta) p(X_k|\mathcal{Y}_k, \theta) dX_k$
5. Marginalize: $\mathcal{L}_{k+1}(\theta; \mathcal{Y}_{k+1}) = \int p(y_{k+1}|X_{k+1}, \theta) p(X_{k+1}|\mathcal{Y}_k, \theta) dX_{k+1}$
6. Update: $p(X_{k+1}|\mathcal{Y}_{k+1}, \theta) = \dfrac{p(y_{k+1}|X_{k+1}, \theta) p(X_{k+1}|\mathcal{Y}_k, \theta)}{p(y_{k+1}|\mathcal{Y}_k, \theta)}$
7. **end for**
8. Accept θ with Metropolis-Hastings probability; otherwise reject
9. **end for**

Kalman Filter / Bayesian Filter

MCMC

Specialization for Linear Systems

Regularization derived from first principles

- Let the state be distributed normally as $X_k \sim \mathcal{N}(m_k, P_k)$
- The negative log-likelihood is equivalent to a time-varying weighted least-squares objective with regularization

$$\mathcal{L}(\theta; \mathcal{Y}_n) = \prod_{k=1}^{n} \mathcal{N}(y_k;\ H(\theta) m_k^-(\theta), S_k)$$

$$-\log \mathcal{L}(\theta; \mathcal{Y}_n) \propto \underbrace{\sum_{k=1}^{n} \|y_k - H(\theta) m_k^-(\theta)\|^2_{S_k^{-1}(\theta)}}_{\text{Low output error when } |S_k| \text{ small}} + \underbrace{\log|2\pi S_k(\theta)|}_{\text{Low output variance}}$$

Where

$$P_k^-(\theta) = A(\theta) P_{k-1}^+(\theta) A^T(\theta) + \Sigma(\theta)$$
$$S_k(\theta) = H(\theta) P_k^-(\theta) H^T(\theta) + \Gamma(\theta)$$

A dynamics matrix
H observation matrix

Outline

Hamiltonian Systems

- In mechanical systems, the Hamiltonian $\mathcal{H}$ is the sum of potential energy U and kinetic energy T

$$\mathcal{H}(q,p) = T(q,p) + U(q,p)$$

q generalized position
p generalized momentum

- Equations of motion are derived from the Hamiltonian

$$\dot{q} = \frac{\partial \mathcal{H}}{\partial p} \qquad \dot{p} = -\frac{\partial \mathcal{H}}{\partial q}$$

- Hamiltonian systems have a number of physical properties
  - Conservation
  - Reversibility
  - Symplecticness

Encoding Symplectic Hamiltonian Systems

Ensures the learned system is Hamiltonian

$$\mathcal{H}(q,p,\theta_\Psi) = \frac{1}{2}p^T p + U(q,\theta_\Psi)$$

Differentiation

$$\dot{q} = p, \qquad \dot{p} = -\frac{\partial U(q,\theta_\Psi)}{\partial q}$$

Leapfrog Method

Conserves Hamiltonian and preserves symplectic structure throughout evaluation

$$\Psi(q_k,p_k;\theta_\Psi) = \begin{bmatrix} q_k + \Delta t p_k - \frac{\Delta t^2}{2}\frac{\partial U(q,\theta_\Psi)}{\partial q}\Big|_{q_k} \\ p_k - \frac{\Delta t}{2}\left(\frac{\partial U(q,\theta_\Psi)}{\partial q}\Big|_{q_k} + \frac{\partial U(q,\theta_\Psi)}{\partial q}\Big|_{q_{k+1}}\right) \end{bmatrix}$$

Non-Separable Systems: Explicit Symplectic Integrator

M. Tao, "Explicit symplectic approximation of nonseparable Hamiltonians: Algorithm and long time performance," Physical Review E, vol. 94, no. 4, p. 043303, 2016.

Introduce fictitious position $\tilde{\mathbf{q}}$ and fictitious momentum $\tilde{\mathbf{p}}$ and define the augmented Hamiltonian as

$$\bar{H}(\mathbf{q},\mathbf{p},\tilde{\mathbf{q}},\tilde{\mathbf{p}}) = \underbrace{H(\mathbf{q},\tilde{\mathbf{p}})}_{H_a} + \underbrace{H(\tilde{\mathbf{q}},\mathbf{p})}_{H_b} + \underbrace{\omega\left(\frac{1}{2}\|\mathbf{q}-\tilde{\mathbf{q}}\|_2^2 + \frac{1}{2}\|\mathbf{p}-\tilde{\mathbf{p}}\|_2^2\right)}_{H_c}$$

Now, $\mathbf{q}$ and $\mathbf{p}$ are decoupled and an explicit symplectic integrator can be defined as

$$\psi^{\Delta t} \coloneqq \psi_{H_a}^{\Delta t/2} \circ \psi_{H_b}^{\Delta t/2} \circ \psi_{H_c}^{\Delta t} \circ \psi_{H_b}^{\Delta t/2} \circ \psi_{H_a}^{\Delta t/2}$$

Where

$$\psi_{H_a}^{\Delta t}\colon \begin{bmatrix}\mathbf{q}\\ \mathbf{p}\\ \tilde{\mathbf{q}}\\ \tilde{\mathbf{p}}\end{bmatrix} \to \begin{bmatrix}\mathbf{q}\\ \mathbf{p}-\Delta t H_{\mathbf{q}}(\mathbf{q},\tilde{\mathbf{p}})\\ \tilde{\mathbf{q}}+\Delta t H_{\tilde{\mathbf{p}}}(\mathbf{q},\tilde{\mathbf{p}})\\ \tilde{\mathbf{p}}\end{bmatrix};\ \psi_{H_b}^{\Delta t}\colon \begin{bmatrix}\mathbf{q}\\ \mathbf{p}\\ \tilde{\mathbf{q}}\\ \tilde{\mathbf{p}}\end{bmatrix} \to \begin{bmatrix}\mathbf{q}+\Delta t H_{\boldsymbol{p}}(\tilde{\mathbf{q}},\mathbf{p})\\ \mathbf{p}\\ \tilde{\mathbf{q}}\\ \tilde{\mathbf{p}}-\Delta t H_{\tilde{\mathbf{q}}}(\tilde{\mathbf{q}},\mathbf{p})\end{bmatrix};\ \psi_{\omega H_c}^{\Delta t}\colon \begin{bmatrix}\mathbf{q}\\ \mathbf{p}\\ \tilde{\mathbf{q}}\\ \tilde{\mathbf{p}}\end{bmatrix} \to \frac{1}{2}\begin{bmatrix}\begin{pmatrix}\mathbf{q}+\tilde{\mathbf{q}}\\ \mathbf{p}+\tilde{\mathbf{p}}\end{pmatrix} + \mathbf{R}(\Delta t)\begin{pmatrix}\mathbf{q}-\tilde{\mathbf{q}}\\ \mathbf{p}-\tilde{\mathbf{p}}\end{pmatrix}\\ \begin{pmatrix}\mathbf{q}+\tilde{\mathbf{q}}\\ \mathbf{p}+\tilde{\mathbf{p}}\end{pmatrix} - \mathbf{R}(\Delta t)\begin{pmatrix}\mathbf{q}-\tilde{\mathbf{q}}\\ \mathbf{p}-\tilde{\mathbf{p}}\end{pmatrix}\end{bmatrix}$$

$$\mathbf{R}(\Delta t) = \begin{bmatrix}\cos(2\omega\Delta t)\,\mathbf{I} & \sin(2\omega\Delta t)\,\mathbf{I}\\ -\sin(2\omega\Delta t)\,\mathbf{I} & \cos(2\omega\Delta t)\,\mathbf{I}\end{bmatrix}$$

Such that a symplectic approximation of the dynamics is

$$[\mathbf{q}^T\ \ \mathbf{p}^T\ \ \tilde{\mathbf{q}}^T\ \ \tilde{\mathbf{p}}^T]_{k+1} = \psi^{\Delta t}([\mathbf{q}^T\ \ \mathbf{p}^T\ \ \tilde{\mathbf{q}}^T\ \ \tilde{\mathbf{p}}^T]_k)$$

Outline

1. Existing approaches
2. Probabilistic formulation
3. Algorithm/Marginal likelihood
4. Hamiltonian Systems
5. **Results**
6. Takeaways

Symplectic vs. non-symplectic integration during learning

- Methods will sometimes use a non-symplectic integrator during training
  - Greydanus, S., Dzamba, M., & Yosinski, J. (2019). Hamiltonian neural networks. *Advances in neural information processing systems*, *32*.
  - Zhong, Y. D., Dey, B., & Chakraborty, A. (2020). Symplectic ODE-Net: Learning Hamiltonian Dynamics with Control. In *International Conference on Learning Representations*.
- Other works have shown improved results can be achieved with a symplectic integrator
  - Toth, P., Rezende, D. J., Jaegle, A., Racanière, S., Botev, A., & Higgins, I. (2020). Hamiltonian Generative Networks. In *International Conference on Learning Representations*.Z. Chen, J. Zhang, M. Arjovsky, and L. Bottou,
  - Chen, Z., Zhang, J., Arjovsky, M., & Bottou, L. (2020). Symplectic Recurrent Neural Networks. In *International Conference on Learning Representations*.
- However, they compare integrators of differing order accuracy

The following results provide:
1. Comparison using symplectic and non-symplectic integrators of comparable order accuracy
2. Quantification of the uncertainty in each of the estimates

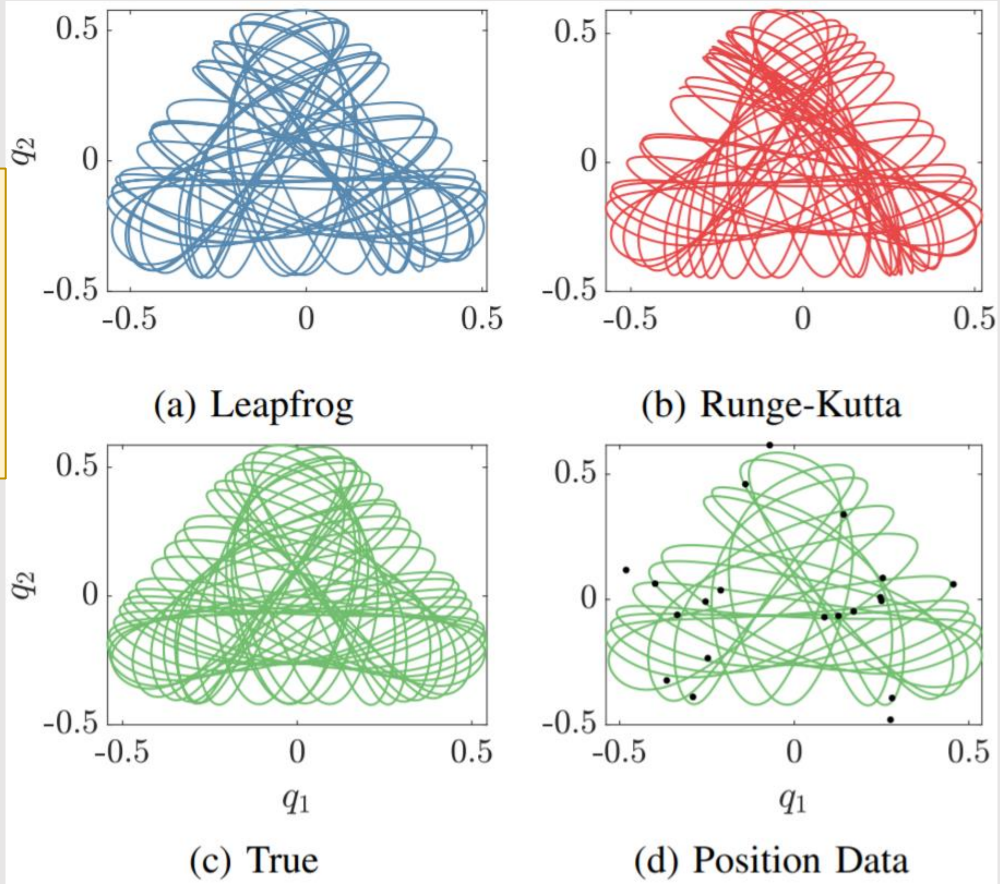
Hénon-Heiles - symplectic vs rk integration

The symplectic approach learns a more accurate Hamiltonian

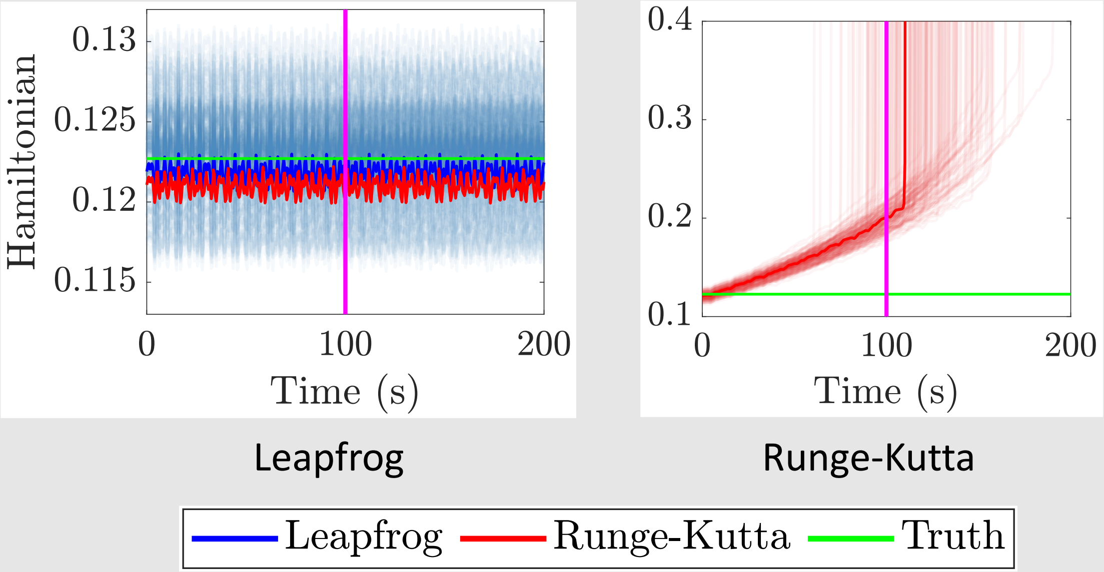
Truth: $U(q_1, q_2) = \frac{1}{2}q_1^2 + \frac{1}{2}q_2^2 + q_1^2 q_2 - \frac{1}{3}q_2^3$

Data Generation:
- $n = 20$
- $\Delta t = 5$
- $\sigma = 0.05$

Phase plots



Hamiltonian over time



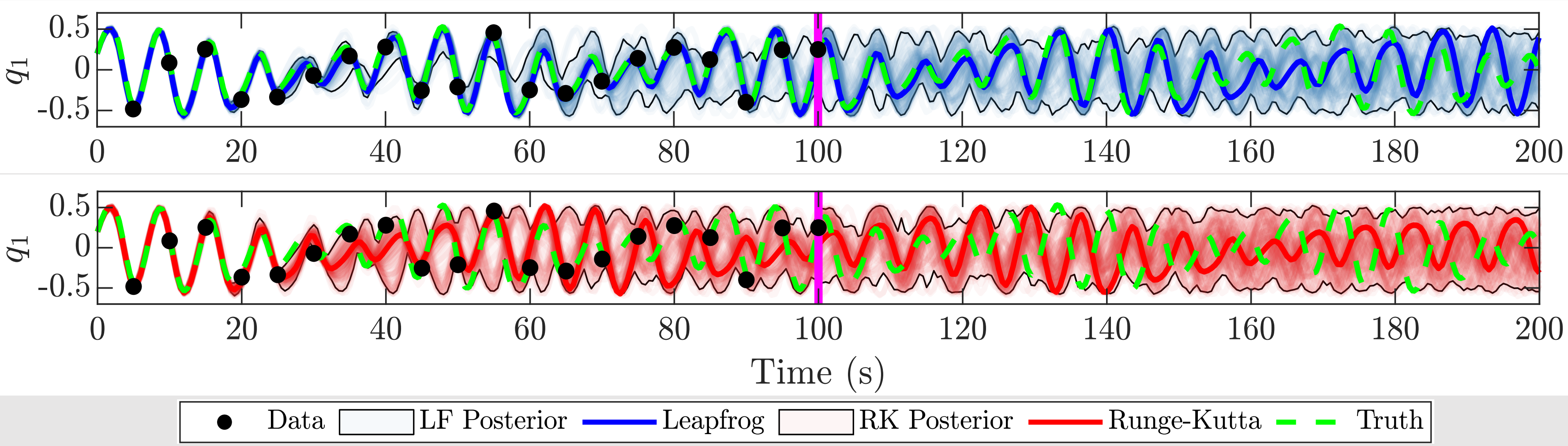
The method equipped with RK must learn a smaller Hamiltonian to compensate for being non-conservative

Relative mean error:
Leapfrog: 0.7%; Runge-Kutta: 1.3%

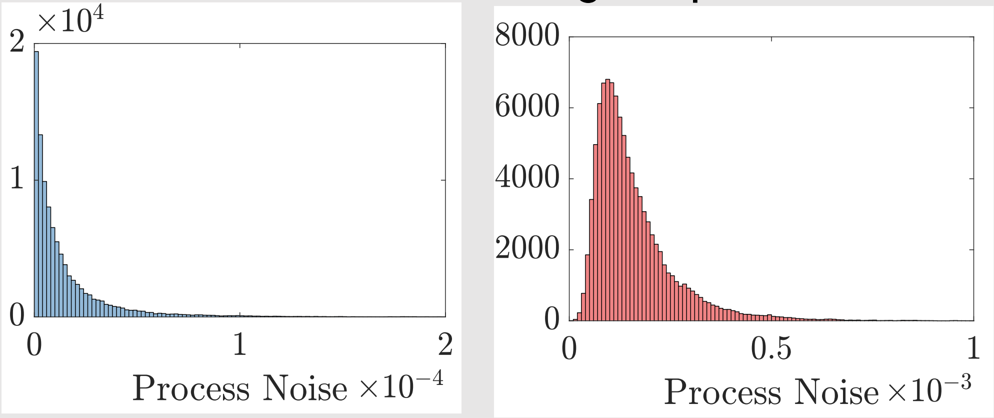
During testing, leapfrog integrator is used on both models

Hénon-Heiles: symplectic approach yields greater certainty

Posterior estimates of q_1 trajectory



Process noise marginal posteriors



Leapfrog   Runge-Kutta

Symplectic approach learns a model with an order of magnitude greater certainty

Nonseparable Hamiltonian: Cherry problem

- System possesses a negative energy mode that causes explosive growth of arbitrarily small perturbations

$$H(q_1, q_2, p_1, p_2) = \frac{1}{2}(q_1^2 + p_1^2) - (q_2^2 + p_2^2) + \frac{1}{2}p_2(p_1^2 - q_1^2) - q_1 q_2 p_1$$

$$\mathbf{y}_k = [\mathbf{q}_k \quad \mathbf{p}_k]^T(1 + u_k), \qquad \text{where } u_k \sim \mathcal{U}[-0.10 \ \ 0.10]$$

- Parametrization: $\Phi(\mathbf{q}, \mathbf{p})$ is vector of Legendre polynomials up to total order 3

$$\widetilde{H}(\mathbf{x}, \theta) = \Phi^T(\mathbf{x})\theta, \text{ where } \mathbf{x} = [q_1 \ q_2 \ p_1 \ p_2]^T$$

- Data generated from five trajectories with random initial conditions

- Training: $\mathbf{x}^{(i)}(0) \sim \mathcal{N}(\mathbf{x}^{test}(0), \ 0.05^2 I_4)$ for $i = 1, \dots, 5$
- Testing: $\mathbf{x}^{test}(0) = [0.15 \ \ 0.10 \ \ -0.05 \ \ 0.10]^T$

$$n = 21$$
$$\Delta t = 0.4$$

Goals: Show the Bayesian algorithm…

1. Provides good estimates even when modeling assumptions are not perfectly met
2. Generalizes well on initial conditions outside training set
3. Outperforms a least-squares algorithm

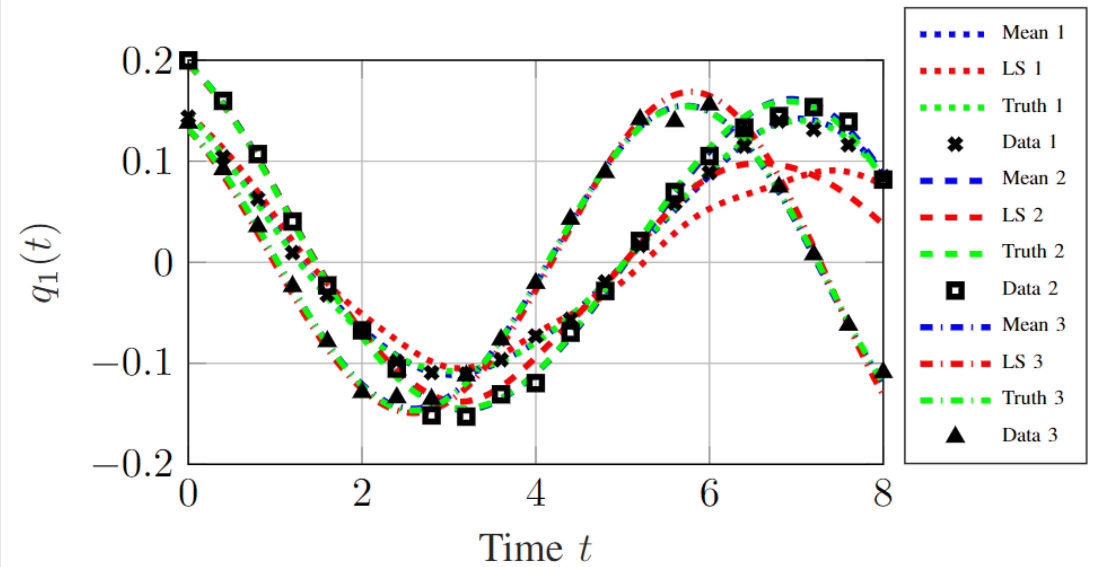
- For learning, an explicit symplectic integrator with integration timestep of 0.01 is used
- We compare the Bayesian posterior to the following least squares (LS) fit[1]:

$$\underset{\theta}{\operatorname{argmin}} \|\nabla\Phi^T(\mathbf{x})\theta - \dot{\mathbf{x}}\|$$

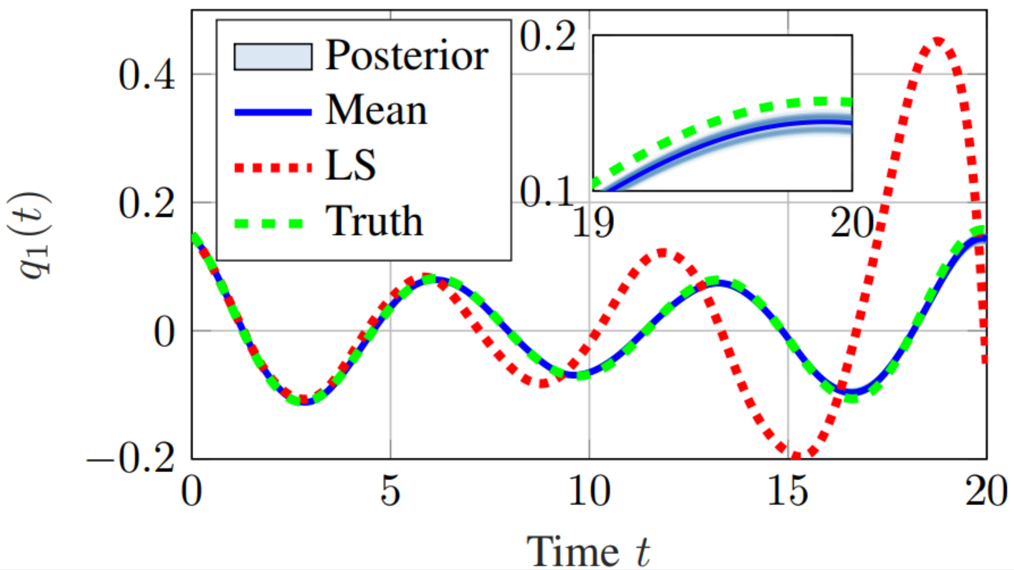
1. Wu, K., Qin, T., & Xiu, D. (2020). Structure-preserving method for reconstructing unknown Hamiltonian systems from trajectory data. *SIAM Journal on Scientific Computing*, *42*(6), A3704-A3729.

Bayesian estimate generalizes well outside of training data

Subset of training



Testing



Relative error: $e(t_k) = \frac{\|\hat{\mathbf{x}}_{1:k} - \mathbf{x}_{1:k}\|_F}{\|\mathbf{x}_{1:k}\|_F}$

Length of time where $e(t) < 10\%$

Least squares	**Mean**
$t = 1.49$	$t = 18.22$

Conclusions

- Optimally accounting for different types of uncertainty can lead to robustness even for chaotic systems
- Modeling deterministic systems with stochastic models introduces built-in regularization and optimization benefits
- Conservation laws can be encoded through integration with appropriate symplectic integrators

Read more

1. Galioto, N., & Gorodetsky, A. A. (2020). Bayesian system ID: optimal management of parameter, model, and measurement uncertainty. *Nonlinear Dynamics*, *102*(1), 241-267.
2. Galioto, N., & Gorodetsky, A. A. (2020) "Bayesian identification of Hamiltonian dynamics from symplectic data." *2020 59th IEEE Conference on Decision and Control (CDC)*. IEEE.
3. Sharma, H., Galioto, N., Gorodetsky, A. A., & Kramer, B. (2022). Bayesian Identification of Nonseparable Hamiltonian Systems Using Stochastic Dynamic Models. *arXiv preprint arXiv:2209.07646*.

Funding

AFOSR Program in Computational Mathematics