Bayesian Identification of Nonseparable Hamiltonian Systems Using Stochastic Dynamic Models

Harsh Sharma[1*], Nicholas Galioto[2*], Alex A. Gorodetsky[2], and Boris Kramer[1]

Abstract— This paper proposes a probabilistic Bayesian formulation for system identification (ID) and estimation of nonseparable Hamiltonian systems using stochastic dynamic models. Nonseparable Hamiltonian systems arise in models from diverse science and engineering applications such as astrophysics, robotics, vortex dynamics, charged particle dynamics, and quantum mechanics. The numerical experiments demonstrate that the proposed method recovers dynamical systems with higher accuracy and reduced predictive uncertainty compared to state-of-the-art approaches. The results further show that accurate predictions far outside the training time interval in the presence of sparse and noisy measurements are possible, which lends robustness and generalizability to the proposed approach. A quantitative benefit is prediction accuracy with less than 10% relative error for more than 12 times longer than a comparable least-squares-based method on a benchmark problem.

I. INTRODUCTION

Nonseparable Hamiltonian systems arise as models in many science and engineering applications such as multibody dynamics and control in robotics [1], the Kozai-Lidov mechanism in astrophysics [2], particle accelerators in accelerator physics [3], 3D vortex dynamics in fluid mechanics [4], and the nonlinear Schrödinger equation in quantum mechanics [5]. These systems demonstrate complex nonlinear behavior while possessing an underlying highly structured geometry encoded by a Hamiltonian. Uncovering a system's Hamiltonian can reveal key insights into its physical properties such as mass or energy conservation. Learning Hamiltonian models directly from data is becoming increasingly important in diverse areas such as astrophysics, robotics, fluid dynamics, plasma physics, and quantum mechanics where first-principle modeling can yield highly complex model structures or such models are not available.

For these purposes, many recent methods have embedded strong physics-motivated inductive priors into their learning framework to develop structure-preserving neural networks for Hamiltonian systems, e.g., [6], [7], [8], [9]. In another research direction, techniques based on sparse identification of nonlinear dynamics (SINDy) [10] and orthogonal polynomials [11] have also been used for learning Hamiltonian systems from data, but these techniques tend to break down when the data are noisy/sparse since they rely on numerical approximations of the time derivatives of the data. To handle uncertainty and increase robustness, Bayesian inference techniques based on Gaussian process regression [12], [13] and/or consideration of stochastic dynamics [14] have been developed. However, a majority of these approaches assume that the Hamiltonian system is separable, i.e., the system Hamiltonian can be written as $H(\mathbf{q}, \mathbf{p}) = T(\mathbf{p}) + U(\mathbf{q})$ where $\mathbf{q}$ is the position, $\mathbf{p}$ is the momentum, $T(\mathbf{p})$ is the kinetic energy, and $U(\mathbf{q})$ is the potential energy.

For nonseparable Hamiltonian systems, nonseparable symplectic neural networks were recently developed in [15] by embedding an appropriate symplectic integrator into the neural network architecture. Instead of learning the symplectic maps, generating function neural networks (GFNNs) in [16] learn Hamiltonian models by approximating the generating functions corresponding to these symplectic maps. Although both of these approaches learn nonseparable Hamiltonian systems from noisy data, they use an optimization objective and modeling format that requires large datasets. These methods use data from over a thousand (up to a million in the case of GFNN) short trajectories, each consisting of five or fewer data points, even for one-dimensional problems.

In this paper, we learn a nonseparable Hamiltonian system from noisy and sparse data using a stochastic model to account for model errors that always exist in any approximation format. Previously, we have shown the consideration of such stochastic process noise aids system recovery by smoothing the learning objective [17]. Building on the Bayesian system identification (ID) work in [17], we embed the physics underlying nonseparable Hamiltonian structure into the learning formulation to develop a probabilistic learning method that preserves the symplectic structure intrinsic to Hamiltonian dynamics. The main contributions of this work are:

1) We extend the Bayesian system ID method of [17] to nonseparable Hamiltonian systems by embedding an explicit structure-preserving numerical integrator within the physics-informed objective;

2) We present detailed numerical results for direct comparison between symplectic and non-symplectic approaches that demonstrate the advantage of symplectic structure preservation. The symplectic approach achieves 30% reduction in relative state error over the non-symplectic approach;

3) We apply the Bayesian method to a dataset with multiple trajectories and non-Gaussian measurement noise for the first time to demonstrate its versatility and robustness. Through this study, we also show that the Gaussian filtering approaches we leverage are not

*These authors contributed equally.

[1]H. Sharma and B. Kramer are with the Department of Mechanical and Aerospace Engineering, University of California San Diego, San Diego, California, USA. {hasharma, bmkramer}@ucsd.edu

[2]N. Galioto and A.A. Gorodetsky are with the Department of Aerospace Engineering, University of Michigan, Ann Arbor, Michigan, USA. {ngalioto, goroda}@umich.edu

overly restrictive even for nonlinear dynamics. After training on a noisy dataset, the proposed method has a relative error below 10% and does so for more than 12 times longer than the least-squares comparison method on a trajectory outside the training set.

This paper is structured as follows. Section II reviews the basics of nonseparable Hamiltonian systems and describes a probabilistic formulation of the system ID problem. Section III presents the proposed structure-preserving algorithm for learning nonseparable Hamiltonian systems. In Section IV, we apply the proposed Bayesian algorithm to two datasets generated from nonseparable Hamiltonian systems. We also compare results from the Bayesian approach with other structure-preserving learning works. Finally, in Section V, we provide concluding remarks and future research directions.

II. BACKGROUND

In Section II-A, we introduce nonseparable Hamiltonian systems, followed by their structure-preserving time integration using explicit symplectic integrators in Section II-B. We then review the probabilistic modeling framework in Section II-C. This provides the necessary background for the structure-preserving Bayesian learning method described in Section III.

A. Nonseparable Hamiltonian systems

The governing equations for finite-dimensional canonical Hamiltonian systems are

$$\dot{\mathbf{q}} = \frac{\partial H(\mathbf{q}, \mathbf{p})}{\partial \mathbf{p}}, \qquad \dot{\mathbf{p}} = -\frac{\partial H(\mathbf{q}, \mathbf{p})}{\partial \mathbf{q}}, \tag{1}$$

where H, the Hamiltonian, is a function of the canonical position $\mathbf{q}$ and momentum $\mathbf{p}$. Many Hamiltonian systems of interest to engineers and scientists (e.g., [1]-[5]) are not additively separable with respect to functions of the position $\mathbf{q}$ and momentum $\mathbf{p}$. Such Hamiltonian systems are said to be *nonseparable*. Unlike the separable Hamiltonian systems with $H(\mathbf{q}, \mathbf{p}) = T(\mathbf{p}) + U(\mathbf{q})$, the governing equations in (1) cannot be further simplified for nonseparable Hamiltonians.

These governing equations possess physically meaningful geometric properties that can be described in the form of symmetries, symplecticity, first integrals, and energy conservation. Preservation of these qualitative features in a numerical simulation is crucial for accurate long-time prediction of Hamiltonian dynamics. Using ideas from geometric mechanics, the field of geometric numerical integration has developed a variety of structure-preserving time integrators for Lagrangian/Hamiltonian systems, e.g., [18], [19].

Example 1 (Cherry Problem): Consider the following nonseparable Hamiltonian from [20] with

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

This four-dimensional dynamical system is a challenging example because it possesses a negative energy mode (NEM) that leads to an explosive nonlinear growth of perturbations

for arbitrarily small disturbances. These NEMs occur in several important infinite-dimensional dynamical systems, e.g., gravitational instability of interpenetrating fluids [21] and magnetosonic waves in the solar atmosphere [22].

B. Tao's explicit symplectic integrator

While explicit symplectic integration has been extensively studied for separable Hamiltonian systems and specific subclasses of nonseparable Hamiltonian systems, explicit symplectic approximations of general nonseparable Hamiltonian systems $H(\mathbf{q}, \mathbf{p})$ is an active research topic. Recently, explicit symplectic integrators for arbitrary nonseparable Hamiltonian systems were presented in [23]. The derivation of these integrators is based on the idea of an extended phase space. For an arbitrary nonseparable Hamiltonian $H(\mathbf{q}, \mathbf{p})$, we first introduce fictitious configuration $\tilde{\mathbf{q}}$ and fictitious momentum $\tilde{\mathbf{p}}$ corresponding to $\mathbf{q}$ and $\mathbf{p}$, respectively. Next, we define an augmented Hamiltonian

$$\bar{H}(\mathbf{q}, \mathbf{p}, \tilde{\mathbf{q}}, \tilde{\mathbf{p}}) := \underbrace{H(\mathbf{q}, \tilde{\mathbf{p}})}_{H_a} + \underbrace{H(\tilde{\mathbf{q}}, \mathbf{p})}_{H_b}$$
$$+ \omega \cdot \underbrace{\left(\|\mathbf{q} - \tilde{\mathbf{q}}\|_2^2/2 + \|\mathbf{p} - \tilde{\mathbf{p}}\|_2^2/2 \right)}_{H_c}, \tag{3}$$

where $H_a := H(\mathbf{q}, \tilde{\mathbf{p}})$ and $H_b := H(\tilde{\mathbf{q}}, \mathbf{p})$ correspond to two copies of the original nonseparable Hamiltonian system with mixed-up positions and momenta; H_c is an artificial restraint; and ω is a constant that controls the binding of the two copies. The governing equations for the augmented Hamiltonian system (3) are

$$\dot{\mathbf{q}} = \nabla_{\mathbf{p}} H(\tilde{\mathbf{q}}, \mathbf{p}) + \omega(\mathbf{p} - \tilde{\mathbf{p}}), \quad \dot{\mathbf{p}} = -\nabla_{\mathbf{q}} H(\mathbf{q}, \tilde{\mathbf{p}}) - \omega(\mathbf{q} - \tilde{\mathbf{q}}),$$
$$\dot{\tilde{\mathbf{q}}} = \nabla_{\tilde{\mathbf{p}}} H(\mathbf{q}, \tilde{\mathbf{p}}) + \omega(\tilde{\mathbf{p}} - \mathbf{p}), \quad \dot{\tilde{\mathbf{p}}} = -\nabla_{\tilde{\mathbf{q}}} H(\tilde{\mathbf{q}}, \mathbf{p}) - \omega(\tilde{\mathbf{q}} - \mathbf{q}).$$

The introduction of fictional variables $\tilde{\mathbf{q}}$ and $\tilde{\mathbf{p}}$ along with a specific choice for the augmented Hamiltonian in (3) decouples the position $\mathbf{q}$ and $\mathbf{p}$ in the extended phase space, i.e., $\dot{\mathbf{q}}$ and $\dot{\mathbf{p}}$ are independent of $\mathbf{q}$ and $\mathbf{p}$, respectively. Unlike the original nonseparable Hamiltonian $H(\mathbf{q}, \mathbf{p})$, the augmented Hamiltonian $\bar{H}(\mathbf{q}, \mathbf{p}, \tilde{\mathbf{q}}, , \tilde{\mathbf{p}})$ is amenable to explicit symplectic integration. In this work, we use second-order explicit symplectic method $\psi^{\Delta t}$ based on Strang splitting

$$\psi^{\Delta t} := \psi_{H_a}^{\Delta t/2} \circ \psi_{H_b}^{\Delta t/2} \circ \psi_{\omega H_c}^{\Delta t} \circ \psi_{H_b}^{\Delta t/2} \circ \psi_{H_a}^{\Delta t/2}, \tag{4}$$

where $\psi_{H_a}^{\Delta t}, \psi_{H_b}^{\Delta t}$, and $\psi_{\omega H_c}^{\Delta t}$ are the time-Δt flow of H_a, H_b, and ωH_c. This allows us to obtain an explicit symplectic integrator for the augmented Hamiltonian $\bar{H}$ via composition of explicit symplectic Euler substeps with step size $\Delta t/2$. Explicit update equations for these individual flows can be

written as

$$\psi_{H_a}^{\Delta t} : \begin{bmatrix} \mathbf{q} \\ \mathbf{p} \\ \tilde{\mathbf{q}} \\ \tilde{\mathbf{p}} \end{bmatrix} \rightarrow \begin{bmatrix} \mathbf{q} \\ \mathbf{p} - \Delta t H_{\mathbf{q}}(\mathbf{q}, \tilde{\mathbf{p}}) \\ \tilde{\mathbf{q}} + \Delta t H_{\tilde{\mathbf{p}}}(\mathbf{q}, \tilde{\mathbf{p}}) \\ \tilde{\mathbf{p}} \end{bmatrix},$$

$$\psi_{H_b}^{\Delta t} : \begin{bmatrix} \mathbf{q} \\ \mathbf{p} \\ \tilde{\mathbf{q}} \\ \tilde{\mathbf{p}} \end{bmatrix} \rightarrow \begin{bmatrix} \mathbf{q} + \Delta t H_{\mathbf{p}}(\tilde{\mathbf{q}}, \mathbf{p}) \\ \mathbf{p} \\ \tilde{\mathbf{q}} \\ \tilde{\mathbf{p}} - \Delta t H_{\tilde{\mathbf{q}}}(\tilde{\mathbf{q}}, \mathbf{p}) \end{bmatrix},$$

$$\psi_{\omega H_c}^{\Delta t} : \begin{bmatrix} \mathbf{q} \\ \mathbf{p} \\ \tilde{\mathbf{q}} \\ \tilde{\mathbf{p}} \end{bmatrix} \rightarrow \frac{1}{2} \begin{bmatrix} \begin{pmatrix} \mathbf{q} + \tilde{\mathbf{q}} \\ \mathbf{p} + \tilde{\mathbf{p}} \end{pmatrix} + \mathbf{R}(\Delta t) \begin{pmatrix} \mathbf{q} - \tilde{\mathbf{q}} \\ \mathbf{p} - \tilde{\mathbf{p}} \end{pmatrix} \\ \begin{pmatrix} \mathbf{q} + \tilde{\mathbf{q}} \\ \mathbf{p} + \tilde{\mathbf{p}} \end{pmatrix} - \mathbf{R}(\Delta t) \begin{pmatrix} \mathbf{q} - \tilde{\mathbf{q}} \\ \mathbf{p} - \tilde{\mathbf{p}} \end{pmatrix} \end{bmatrix},$$

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

Given a nonseparable Hamiltonian $H(\mathbf{q}, \mathbf{p})$ with initial condition $[\mathbf{q}(0)^\top, \mathbf{p}(0)^\top]^\top = [\mathbf{q}_0^\top, \mathbf{p}_0^\top]^\top$, we obtain explicit symplectic approximations of the dynamics by integrating the augmented Hamiltonian $\bar{H}(\mathbf{q}, \mathbf{p}, \tilde{\mathbf{q}}, \tilde{\mathbf{p}})$ with $[\mathbf{q}(0)^\top, \mathbf{p}(0)^\top, \tilde{\mathbf{q}}(0)^\top, \tilde{\mathbf{p}}(0)^\top]^\top = [\mathbf{q}_0^\top, \mathbf{p}_0^\top, \mathbf{q}_0^\top, \mathbf{p}_0^\top]^\top$.

C. Bayesian system identification

In the Bayesian learning framework, the Hamiltonians are unknown. Furthermore, they are parametrized in some approximation format, e.g., polynomial or neural network expansion.

We first model the system probabilistically as a hidden Markov model. Let $\mathbf{X}_k = [\mathbf{q}_k^\top, \mathbf{p}_k^\top]^\top \in \mathbb{R}^{2d}$ be the random variable representing a belief of the system's state, and let the subscript k be the index corresponding to the time t_k at which output data $\mathbf{y}_k \in \mathbb{R}^{d_y}$ is observed. The output data is modeled as a realization of the random variable $\mathbf{Y}_k \in \mathbb{R}^{d_y}$ that represents a belief of the systems true output. The state-transition operator $\Psi : \mathbb{R}^{2d} \times \mathbb{R}^{d_{\boldsymbol{\theta}}} \rightarrow \mathbb{R}^{2d}$ maps the state $\mathbf{X}_k$ at an arbitrary time t_k to the state $\mathbf{X}_{k+1}$, and the observation operator $h : \mathbb{R}^{2d} \times \mathbb{R}^{d_{\boldsymbol{\theta}}} \rightarrow \mathbb{R}^{d_y}$ maps the state $\mathbf{X}_k$ to the output $\mathbf{Y}_k$. Both operators may be parametrized by the random variable $\boldsymbol{\theta} \in \mathbb{R}^{d_{\boldsymbol{\theta}}}$. The model of the system dynamics is described by

$$\begin{aligned} \mathbf{X}_{k+1} &= \Psi(\mathbf{X}_k, \boldsymbol{\theta}_\Psi) + \boldsymbol{\xi}_k, & \boldsymbol{\xi}_k &\sim \mathcal{N}(\mathbf{0}, \Sigma(\boldsymbol{\theta}_\Sigma)), \\ \mathbf{Y}_k &= h(\mathbf{X}_k, \boldsymbol{\theta}_h) + \boldsymbol{\eta}_k, & \boldsymbol{\eta}_k &\sim \mathcal{N}(\mathbf{0}, \Gamma(\boldsymbol{\theta}_\Gamma)). \end{aligned} \quad (5)$$

The parameters are partitioned as $\boldsymbol{\theta} = (\boldsymbol{\theta}_\Psi, \boldsymbol{\theta}_h, \boldsymbol{\theta}_\Sigma, \boldsymbol{\theta}_\Gamma)$; the process noise $\boldsymbol{\xi}_k$ and measurement noise $\boldsymbol{\eta}_k$ are additive, zero-mean Gaussian with unknown covariances $\Sigma(\boldsymbol{\theta}_\Sigma)$ and $\Gamma(\boldsymbol{\theta}_\Gamma)$, respectively. Gaussian noise is chosen not only because it is easy to work with, but also because for a given mean and covariance, the Gaussian distribution satisfies the Principle of Maximum Entropy [24], meaning it minimizes the amount of information that must be assumed beyond the first two moments.

In this model, we assume that there exists error/noise in both the state propagation and observation operators, and we therefore learn a stochastic system rather than a deterministic one. The source of the measurement error is typically sensor noise, and the term $\boldsymbol{\eta}_k$ is included in nearly all system ID

models. The error in the state propagation operator arises due to the uncertainty in both the parameters $\boldsymbol{\theta}_\Psi$ and in the choice of the model structure. In other words, unless the model is perfectly known, the output of Ψ will always carry some amount of error with respect to the "true" state. The process noise $\boldsymbol{\xi}_k$ is not usually included in system ID models, but in [17], it was shown to improve the robustness of system ID methods when faced with sparse and/or noisy data. In this paper, we will further demonstrate the benefits of accounting for the process error.

The goal of Bayesian inference is to compute a posterior distribution for $\boldsymbol{\theta}$ that represents available knowledge of the parameter values given all of the data that we have so far collected. This posterior distribution is denoted as $\pi(\boldsymbol{\theta}|\mathcal{Y}_n)$ where $\mathcal{Y}_n := (\mathbf{Y}_1 = \mathbf{y}_1, \ldots, \mathbf{Y}_n = \mathbf{y}_n)$ is the collection of data points $\mathbf{y}_k$ treated as realizations of the random variables $\mathbf{Y}_k$ for $k = 1, \ldots, n$. Representing available knowledge as a distribution rather than a single estimated point is especially desirable in cases where the data are incomplete/weakly informative, such as sparse and/or noisy data. In these cases, the precision of the parameter estimate is limited, and it is therefore useful to be able to represent both the parameters and the predicted system output given these parameters as distributions.

To compute the posterior distribution, we factorize $\pi(\boldsymbol{\theta}|\mathcal{Y}_n)$ using Bayes' theorem:

$$\pi(\boldsymbol{\theta}|\mathcal{Y}_n) = \frac{\mathcal{L}(\boldsymbol{\theta}; \mathcal{Y}_n)\pi(\boldsymbol{\theta})}{\pi(\mathcal{Y}_n)} \propto \mathcal{L}(\boldsymbol{\theta}; \mathcal{Y}_n)\pi(\boldsymbol{\theta}), \quad (6)$$

where $\mathcal{L}(\boldsymbol{\theta}; \mathcal{Y}_n) := \pi(\mathcal{Y}_n|\boldsymbol{\theta})$ is the likelihood, $\pi(\boldsymbol{\theta})$ is the prior, and $\pi(\mathcal{Y}_n)$ is the evidence. Because $\pi(\mathcal{Y}_n)$ does not depend on the parameters, we treat it as a normalizing constant and do not need to evaluate it. The prior is chosen by the user, and is usually chosen to be easy to evaluate. The main challenge of evaluating the posterior then is the evaluation of the likelihood. Since the dynamics model is stochastic, the state $\mathbf{X}_k$ is also uncertain, and, although we are interested in the marginal likelihood $\mathcal{L}(\boldsymbol{\theta}; \mathcal{Y}_n)$, we are only able to directly compute the joint likelihood $\mathcal{L}(\mathcal{X}_n, \boldsymbol{\theta}; \mathcal{Y}_n)$, where $\mathcal{X}_n := (\mathbf{X}_1, \ldots, \mathbf{X}_n)$. To extract the marginal likelihood from the joint likelihood requires the evaluation of the high-dimensional integral $\int \mathcal{L}(\mathcal{X}_n, \boldsymbol{\theta}; \mathcal{Y}_n) \mathrm{d}\mathcal{X}_n$. This costly computation can be avoided, however, by factorizing the marginal likelihood as follows:

$$\mathcal{L}(\boldsymbol{\theta}; \mathcal{Y}_n) = \prod_{k=1}^{n} \mathcal{L}_k(\boldsymbol{\theta}; \mathcal{Y}_k), \quad (7)$$

and by using the algorithm in [25, Thm 12.1] to recursively evaluate each term $\mathcal{L}_k(\boldsymbol{\theta}; \mathcal{Y}_n)$. When there exist nonlinearities in either of the functions $\Psi(\cdot, \cdot)$ or $h(\cdot, \cdot)$, the distributions $\mathcal{L}_k(\boldsymbol{\theta}; \mathcal{Y}_k)$ will be non-Gaussian, but we approximate them as Gaussian using the unscented Kalman filter (UKF) for computational efficiency. See [26], [27], [28] for a similar approach. The computational complexity of the integrator (4) is $\mathcal{O}(d^3 + dG)$, where G denotes the cost of evaluating the gradient of the Hamiltonian. Based on the analysis in [17], the computational complexity of evaluating the marginal

likelihood embedded with Tao's integrator is $\mathcal{O}(nd^3 + ndG)$. For more details on this algorithm, see [17], [14].

For other types of noise in system (5) such as multiplicative or non-Gaussian noise, different filtering algorithms can be used to evaluate the marginal likelihood with possibly improved approximation quality. If the noise is multiplicative, a Kalman filter variation designed for multiplicative noise [29] can be used. If extreme non-Gaussianity is present in the estimation distribution, a particle filter [30] can be used to approximate the marginal likelihood with an empirical discrete distribution rather than a Gaussian, but the cost of the algorithm increases significantly. In Section IV-B, we break the assumption of additive, Gaussian noise and show that the Bayesian algorithm with the UKF can still deliver accurate estimates. The numerical results in Section IV-B demonstrate that the Bayesian algorithm has predictive power even when the assumptions of system (5) are not strictly met.

III. PROBABILISTIC LEARNING OF NONSEPARABLE HAMILTONIAN SYSTEMS

This section presents a novel methodology for incorporating physical knowledge of nonseparable Hamiltonian systems into the Bayesian learning framework. This discussion includes the parametrization of the model dynamics in Section III-A, the selection of a structure-preserving time integrator within the dynamics propagator in Section III-B, and an outline of the proposed computational procedure in Section III-C.

A. Parametrizing nonseparable Hamiltonian systems

Our goal is to leverage prior knowledge of the underlying system to inform the parametrization of a model. Recall from Section II-A that a Hamiltonian system is defined by a function of the generalized position and momentum known as the Hamiltonian. To leverage this fact, we assume that the generalized coordinate frame is known and attempt to directly learn the Hamiltonian. By estimating a model of the Hamiltonian and deriving the dynamics from this model according to (1), we guarantee that the predicted flow will also be Hamiltonian. This modeling choice makes the predictions more physically meaningful and also serves to restrict the model search space, which can help optimization.

As an example of a parametrization of the Hamiltonian, we consider a linear combination of nonlinear basis functions. In this work, we are concerned with nonseparable Hamiltonians, so we denote the model of the Hamiltonian generally as the differentiable function $\tilde{H}(\mathbf{q}, \mathbf{p}, \boldsymbol{\theta}_\Psi)$. Then we approximate the Hamiltonian up to some additive constant C as

$$\tilde{H}(\mathbf{q}, \mathbf{p}, \boldsymbol{\theta}_\Psi) = \Phi^\top(\mathbf{q}, \mathbf{p})\boldsymbol{\theta}_\Psi + C, \qquad (8)$$

where $\Phi(\mathbf{q}, \mathbf{p}) \in \mathbb{R}^N$ is a vector whose ith component is the differentiable basis function $\phi_i : \mathbb{R}^{2d} \to \mathbb{R}$ for $i = 1, \ldots, N$. The gradient of the Hamiltonian follows

$$\nabla \tilde{H}(\mathbf{q}, \mathbf{p}, \boldsymbol{\theta}_\Psi) = \nabla \Phi^\top(\mathbf{q}, \mathbf{p})\boldsymbol{\theta}_\Psi, \qquad (9)$$

where $\nabla \Phi(\mathbf{q}, \mathbf{p}) \in \mathbb{R}^{N \times 2d}$ is a matrix where the ith column is the gradient $\nabla \phi_i$ with respect to the state $[\mathbf{q}^\top, \mathbf{p}^\top]^\top$.

The learning framework of Section II-C is applicable to any arbitrary model parametrization including neural networks and other nonlinear approximations.

B. Embedding explicit symplectic integrators into the Bayesian learning framework

In this subsection, we incorporate the stochastic nonseparable Hamiltonian into the Bayesian system ID framework. The key idea is to exploit knowledge of nonseparable Hamiltonian systems and their structure-preserving time integrators to inform the design of the state propagator $\Psi(\mathbf{q}_k, \mathbf{p}_k; \boldsymbol{\theta}_\Psi)$ in (5). We apply the second-order explicit symplectic integrator $\psi^{\Delta t}$ from (4) to the parametrized nonseparable Hamiltonian (8) to obtain

$$\psi_{\boldsymbol{\theta}}^{\Delta t} := \psi_{\tilde{H}_a(\boldsymbol{\theta})}^{\Delta t/2} \circ \psi_{\tilde{H}_b(\boldsymbol{\theta})}^{\Delta t/2} \circ \psi_{\omega \tilde{H}_c}^{\Delta t} \circ \psi_{\tilde{H}_b(\boldsymbol{\theta})}^{\Delta t/2} \circ \psi_{\tilde{H}_a(\boldsymbol{\theta})}^{\Delta t/2},$$

where $\tilde{H}_a(\boldsymbol{\theta}) := \tilde{H}(\mathbf{q}, \tilde{\mathbf{p}}, \boldsymbol{\theta})$ and $\tilde{H}_b(\boldsymbol{\theta}) := \tilde{H}(\tilde{\mathbf{q}}, \mathbf{p}, \boldsymbol{\theta})$ correspond to two copies of the parametrized nonseparable Hamiltonian model with mixed-up positions and momenta, and $\omega \tilde{H}_c$ is an artificial restraint that controls the binding of the two parametrized copies. We use $\psi_{\boldsymbol{\theta}}^{\Delta t}$ to encode Hamiltonian structure into the propagator, i.e.,

$$\Psi(\mathbf{q}_k, \mathbf{p}_k; \boldsymbol{\theta}_\Psi) := L^\dagger \circ \psi_{\boldsymbol{\theta}}^{\Delta t} \circ L,$$

where $L : [\mathbf{q}_k^\top, \mathbf{p}_k^\top]^\top \to [\mathbf{q}_k^\top, \mathbf{p}_k^\top, \mathbf{q}_k^\top, \mathbf{p}_k^\top]^\top$ duplicates the state to obtain the augmented state, and $L^\dagger : [\mathbf{q}_k^\top, \mathbf{p}_k^\top, \tilde{\mathbf{q}}_k^\top, \tilde{\mathbf{p}}_k^\top]^\top \to [\mathbf{q}_k^\top, \mathbf{p}_k^\top]^\top$ operates on the augmented state to yield the state in the original phase space. This physics-informed choice of the propagator ensures that the learned model will respect the geometric properties of the true process, i.e., the learned model will be a canonical nonseparable Hamiltonian system.

C. Learning setups

We consider two primary learning setups: (1) data are collected from a single trajectory, and the goal is to predict the future state of this trajectory; and (2) data are collected from multiple trajectories with independent initial conditions, and the goal is to predict the state trajectory at an arbitrary initial condition. The first of these follows straightforwardly from the framework of Section II-C, but the second requires a small adaptation.

Let $\mathcal{Y}_n^{(m)}$ be the collection of n data points measured from the trajectory with initial condition $\boldsymbol{x}_0^{(m)}$. The factorization of the likelihood (7) now becomes $\mathcal{L}(\boldsymbol{\theta}; \mathcal{Y}_n^{(1,\ldots,M)}) = \prod_{m=1}^M \mathcal{L}(\boldsymbol{\theta}; \mathcal{Y}_n^{(m)})$, where each term $\mathcal{L}(\boldsymbol{\theta}; \mathcal{Y}_n^{(m)})$ is evaluated using (7). This likelihood factorization has the added benefit of being easily parallelized for computational efficiency.

The second learning setup is also considered in other works related to learning Hamiltonian systems such as [11], [15]. In this work, we focus specifically on comparing to [11] since it uses the same parametrization (8) that we consider. The method in [11] attempts to learn a linear mapping from a dictionary of basis functions to the time derivatives by solving the following optimization problem:

$$\min_{\mathbf{c} \in \mathbb{R}^N} \|\mathbf{A}\mathbf{c} - \mathbf{b}\|_2, \qquad (10)$$

where each element of $\mathbf{A}$ and $\mathbf{b}$ is defined as

$$a_{ij} = \frac{1}{K} \sum_{k=1}^{K} \Big(\nabla \phi_i(\mathbf{q}_k, \mathbf{p}_k) \cdot \nabla \phi_j(\mathbf{q}_k, \mathbf{p}_k) \Big), \qquad (11)$$

$$b_i = \frac{1}{K} \sum_{k=1}^{K} \Big(\big[\dot{\mathbf{p}}_k^\top \;\; -\dot{\mathbf{q}}_k^\top \big]^\top \cdot \nabla \phi_i(\mathbf{q}_k, \mathbf{p}_k) \Big), \qquad (12)$$

for $1 \leq i, j \leq N$. Typically, data of $\dot{\mathbf{q}}$ and $\dot{\mathbf{p}}$ are not available and must therefore be numerically approximated. Following the approach of [11], we use a second-order finite difference method for this approximation. Since this method solves a linear least-squares problem, we henceforth refer to it as the least-squares (LS) method.

D. Computational procedure

This subsection details the computational procedure that we use to draw samples from the posterior and to make predictions using these samples.

To draw samples, we use Markov chain Monte Carlo (MCMC) sampling. Specifically, we use the delayed rejection adaptive Metropolis (DRAM) [31] MCMC sampler. The proposal is a Gaussian random walk, and the covariance of this Gaussian is scaled by a factor of $\gamma = 0.01$ to use as the second-tier proposal. We begin adaptation of the proposal covariance after 200 samples and add $\varepsilon = 10^{-10}$ to the diagonal to maintain positive semidefiniteness. In all of the examples, we start sampling at the MAP point found using MATLAB's `fmincon` optimizer.

Once we have samples from the posterior, we need to choose how to use them for prediction. Two specific values of the posterior predictive distribution, defined as $\pi(\mathcal{X}_T | \mathcal{Y}_n) = \int \pi(\mathcal{X}_T | \boldsymbol{\theta}) \pi(\boldsymbol{\theta} | \mathcal{Y}_n) \mathrm{d}\boldsymbol{\theta}$ for arbitrary positive integer T, are used in the results. First, the maximum a posteriori (MAP) value $\boldsymbol{\theta}^{\mathrm{MAP}}$ of the parameter posterior is simply its maximum. We use this to obtain the prediction $\boldsymbol{x}_k = \Psi^k(\boldsymbol{x}_0, \boldsymbol{\theta}^{\mathrm{MAP}})$, where Ψ^k denotes k compositions of the state propagator Ψ for any $k \in \mathbb{Z}_+$. The other value is the expected value of the posterior predictive distribution $\boldsymbol{x}_{1:T} = \mathbb{E}[\pi(\mathcal{X}_T | \mathcal{Y}_n)]$. The expected value is estimated as $\boldsymbol{x}_k = \frac{1}{P} \sum_{i=1}^{P} \Psi^k(\boldsymbol{x}_0, \boldsymbol{\theta}_i)$ for any $k \in \mathbb{Z}_+$, where $\boldsymbol{\theta}_i$ is the ith MCMC sample and P the total number of MCMC samples. Typically, we use a subset of samples drawn at regular intervals to perform this estimation for the sake of computational practicality. We will sometimes refer to this prediction point simply as the mean estimate. The decision of which point to use can depend on a number of variables such as the dynamics or the shapes of the parameter posterior and/or posterior predictive distributions. Since the points are easy to compute after sampling, heuristic choices are typically an acceptable decision method.

IV. Numerical Experiments

In this section, we apply the Bayesian learning method presented in Section III to the Cherry problem discussed in Example 1. First, we compare the structure-preserving Bayesian learning approach to a 'physics-ignorant' approach where the propagator is equipped with a non-symplectic integrator of the same order that does not preserve the underlying

Hamiltonian structure. The results of this comparison are provided in Section IV-A and show the improvements in accuracy gained by using a structure-preserving symplectic integrator over a non-symplectic integrator of equivalent order accuracy. Section IV-B, gives a comparison between the proposed method and the LS method described in III-C. This section demonstrates the greater robustness of the proposed approach to noisy data.

In both sections, we parametrize the Hamiltonian with polynomials up to total order three for a total of $N = 34$ basis functions. The type of polynomials used in each example is given in the corresponding section. Each covariance matrix is parametrized as an identity matrix scaled by an unknown parameter: $\Sigma(\boldsymbol{\theta}_\Sigma) = \theta_{35} \mathbf{I}_4$ and $\Gamma(\boldsymbol{\theta}_\Gamma) = \theta_{36} \mathbf{I}_4$. We place a Laplacian prior on $\boldsymbol{\theta}_\Psi$ to promote sparsity and half-normal priors on $\boldsymbol{\theta}_\Sigma$ and $\boldsymbol{\theta}_\Gamma$, which are the distributions of the absolute value of a zero-mean Gaussian random variable. The observation operator h is the identity such that we measure the full state. The data and predictions are both generated using the explicit symplectic integrator using a time step of $\Delta t = 0.01$.

A. Training with a single initial condition

In this study, we consider the case where noisy data from a single initial condition are used to learn the model, and we compare numerical results far outside the training data regime to demonstrate the generalizability of the presented Bayesian learning algorithm.

For training the model, we select the initial condition $[q_1(0), q_2(0), p_1(0), p_2(0)]^\top = [0.15, 0.1, -0.05, 0.1]^\top$ and integrate the nonseparable Hamiltonian system up to $t = 8$. We collect the data on the full state every 0.4 time units for a total of 21 data points with zero-mean Gaussian noise with standard deviation of $\sigma = 0.01$. For the learning problem, we parametrize the nonseparable Hamiltonian using monomial basis functions. To ensure proper convergence, we draw 2×10^5 samples, and discard the first 10^5 samples as a conservative burn-in period.

Both symplectic and non-symplectic approaches in this study use a learning time step of $\Delta t = 0.05$ (five times the time step used for prediction) to reduce computational time. Despite the large learning step, the symplectic approach gives accurate predictions far outside the training data regime, see Figure 2a. For the non-symplectic approach, we use a second-order Runge-Kutta integrator that does not preserve the underlying Hamiltonian structure.

The MAP parameters of each posterior are used to simulate for $t = 16$ (100 % outside the training data regime) using the explicit symplectic integrator. Figure 1 shows the learned Hamiltonian values for the MAP point and their corresponding posteriors for both symplectic and non-symplectic approaches. The true Hamiltonian is -0.78×10^{-2}, and the value of the Hamiltonian learned by the symplectic approach is approximately -0.68×10^{-2}. The Hamiltonian learned by the non-symplectic approach is approximately -0.81×10^{-2}. Compared to the posterior from the symplectic approach in Figure 1a, the posterior from the non-symplectic approach

in Figure 1b reflects greater uncertainty in the Hamiltonian estimate.

Figure 2 compares the reconstructed trajectories of the system's first state using the MAP point and the samples from the posterior. For a fair comparison between the symplectic and non-symplectic approaches, each trajectory is integrated using the explicit symplectic integrator.[1] We see in Figure 2 that the MAP estimate from the symplectic approach aligns closely with the truth for the entire length of the simulation, while the MAP point from the non-symplectic approach gives inaccurate results after the end of the training time interval at $t = 8$.

We also compute the following relative state error for both the training data regime and testing data regime

$$e(i : j) = \frac{\|\hat{\boldsymbol{x}}_{i:j} - \boldsymbol{x}_{i:j}\|_F}{\|\boldsymbol{x}_{i:j}\|_F} \qquad (13)$$

for positive integers $i \leq j$ where $\boldsymbol{x}_{i:j} \in \mathbb{R}^{2d \times (j-i+1)}$ is the true state trajectory at times $t_i, \ldots, t_j$, and $\hat{\boldsymbol{x}}_{i:j}$ is the estimate of this trajectory. In the training data regime ($t = 0$ to $t = 8$), the relative state error, $e(1 : n)$, for the symplectic approach is 0.0877 whereas the relative state error for the non-symplectic approach is 0.1307. In the testing data regime ($t = 8$ to $t = 16$), the relative state error, $e(n + 1 : 2n)$, for the symplectic approach is 0.2173 whereas the relative state error for the non-symplectic approach is 0.3326. Thus, the symplectic approach achieves approximately 30% reduction in state error over the non-symplectic approach.

B. Training with multiple initial conditions

In this subsection, we consider the case where data are collected from a number of different trajectories with independent initial conditions (ICs), and our goal is to estimate a model of the system that can give good prediction for an arbitrary IC.

To gather the training data, we first randomly sample $M = 5$ ICs from a Gaussian centered at the testing IC with standard deviation 0.05. The testing IC is the same IC used in the previous example. For each training IC, we integrate the system for up to $t = 8$ and measure the state every 0.4 time units for a total of 21 data points per trajectory including the IC. Any trajectories that diverged during this time period had their IC resampled. To complete the training dataset, we add 10% relative noise, i.e., $\boldsymbol{Y}_k = \boldsymbol{X}_k(1 + u_k)$ where $u_k \sim \mathcal{U}[-0.10, 0.10]$ and $\mathcal{U}$ is the uniform distribution. This is the same noise form used by [11].

We parametrize the Hamiltonian for the learning problem with Legendre polynomials as basis functions and integrate with the symplectic integrator with $\Delta t = 0.01$ during training. To find the MAP, we use the LS estimate as the optimization starting point for $\boldsymbol{\theta}_\Psi$ and the starting point

---

[1]We assume that a non-symplectic learning approach would still actually integrate the learned model with a symplectic integrator. This setting is similar to the approach considered in [6] or [32] where the authors use non-symplectic integrators for training structure-preserving neural networks and then integrate the learned Hamiltonian models using symplectic integrators.

(a) Symplectic approach
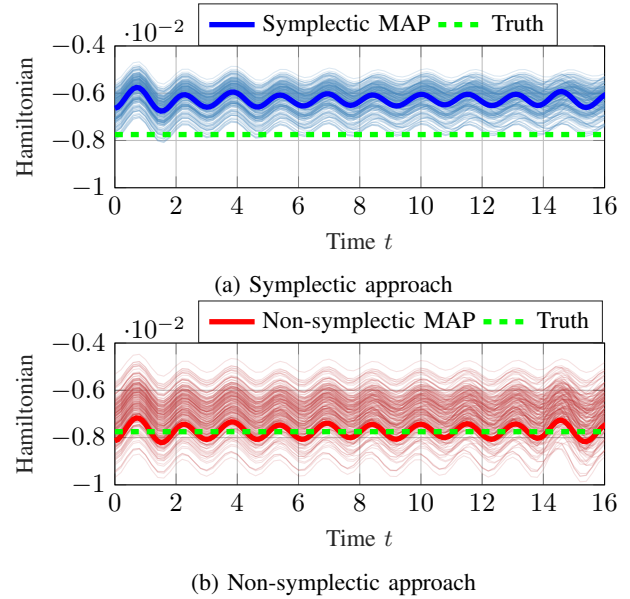
(b) Non-symplectic approach

Fig. 1: The learned Hamiltonians. The lighter blue and red lines in (a) and (b) are the samples from the posteriors from the symplectic and non-symplectic learning approaches. Both approaches estimate the true Hamiltonian accurately with an absolute energy error of less than 10^{-3}. However, the non-symplectic approach has much greater uncertainty.

10^{-6} for variance parameters $\boldsymbol{\theta}_\Sigma$ and $\boldsymbol{\theta}_\Psi$. Then we draw 10^4 samples from the posterior.

We compare the proposed structure-preserving learning method to the LS method and present the results in Figure 3, where every learned model was integrated with the explicit symplectic integrator. Note that although [11] suggested a method for denoising the data, we found that using the raw data gave a better LS estimate in this problem. The training data and ground truth are shown alongside the point estimates from the LS and Bayesian methods in Figure 3a. Even though the training ICs are all near the testing IC, we observe that the testing trajectory in Figure 3b differs significantly from the training trajectories. Figure 3a also shows that the mean estimate is able to fit the training data fairly well, whereas the LS estimate struggles due to the noisiness/sparsity of the data. The fact that the mean estimate can fit the training data even when the measurement noise is non-Gaussian further demonstrates the robustness of the proposed approach.

Next, we illustrate the posterior predictive distribution on the test IC in Figure 3b by plotting the trajectory estimate from every 200th MCMC sample for a total of 50 samples. We see that as time increases, the posterior starts to widen, reflecting the growing uncertainty in the state with time. We also observe that although both the LS and posterior predictive mean match the truth very closely through $t = 4$, the LS estimate deviates from the truth much earlier and much more drastically than the mean estimate. Figure 3b demonstrates that the Bayesian method is able to find models that generalize well outside of the training set of the data.

(a) Symplectic approach
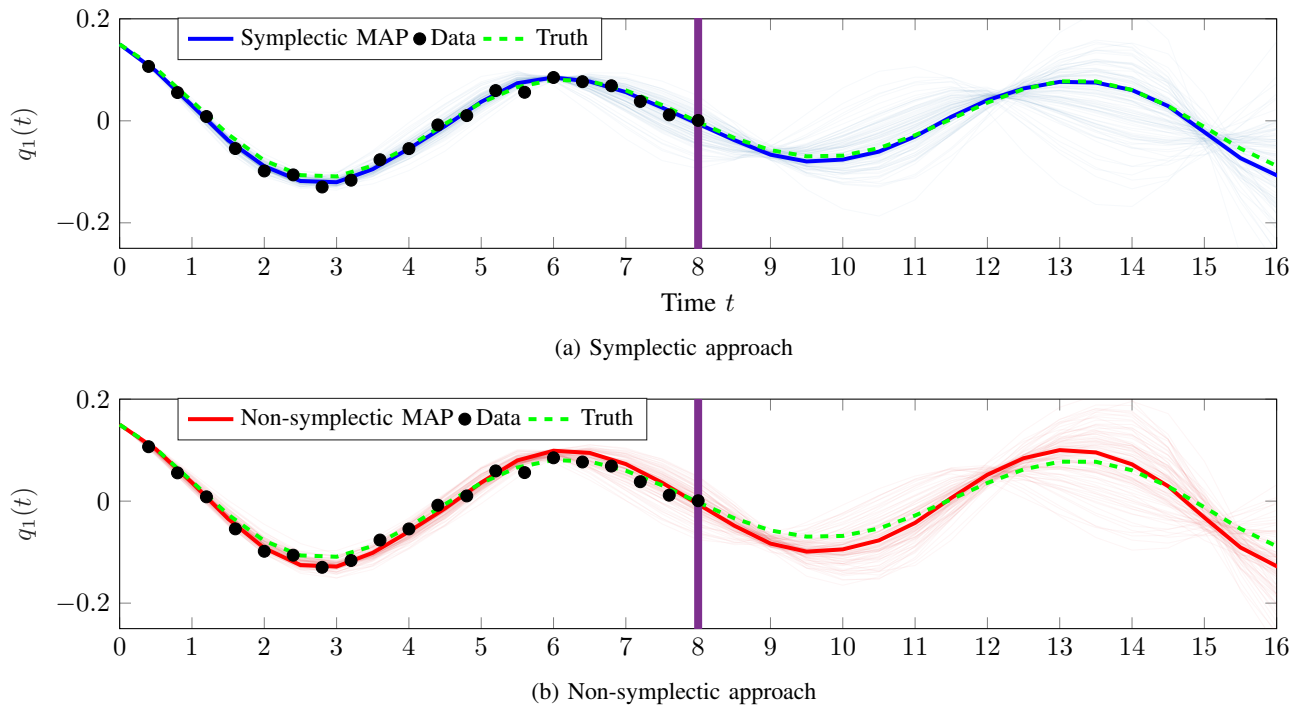

(b) Non-symplectic approach

Fig. 2: Reconstruction error comparison for the training with single IC case. The MAP estimate for the symplectic approach captures the nonseparable Hamiltonian dynamics accurately at $t = 16$ which is 100% outside the training data whereas the MAP estimate for the non-symplectic approach yields inaccurate predictions outside the training data. The posterior for both approaches grows as the system evolves further in time. The purple line indicates the end of the data collection period.

To compare the methods quantitatively, consider the relative error (13) over the first k time steps, i.e., $e(1 : k)$. The relative error of the mean estimate on the test IC stays under 10% for $t = 18.22$, while the LS estimate stays under this threshold for only $t = 1.49$. The Bayesian method therefore yields better long-time prediction performance.

V. CONCLUSIONS

We have developed a structure-preserving Bayesian learning approach for system ID of nonseparable Hamiltonian systems using stochastic dynamic models. This approach expands the Bayesian system ID work in [17] by incorporating knowledge of nonseparable Hamiltonian systems into the learning framework. The numerical results demonstrate the advantage of preserving the underlying geometric structure for a challenging nonseparable Hamiltonian system that possesses a negative energy mode. Hamiltonian models learned using the proposed method provide accurate predictions far outside the training time interval and perform well even for unknown initial conditions.

In future work, we would like to use neural networks for parametrization in the proposed framework for learning non-polynomial Hamiltonian systems. Moreover, while in this work we assume additive noises in the probabilistic model of the system dynamics (5), in many applications the noise is multiplied by the state. It would be desirable to extend the current approach to dynamical systems with multiplicative noises. Finally, we have applied the structure-preserving Bayesian learning method to a system with four states. We would like to combine the proposed approach with Hamiltonian operator inference [33] for structure-preserving learning of reduced-order models of large-scale Hamiltonian models.

VI. ACKNOWLEDGMENTS

REFERENCES

[1] D. Serra, F. Ruggiero, A. Donaire, L. R. Buonocore, V. Lippiello, and B. Siciliano, "Control of nonprehensile planar rolling manipulation: A passivity-based approach," *IEEE Transactions on Robotics*, vol. 35, no. 2, pp. 317–329, 2019.

[2] G. Li, S. Naoz, M. Holman, and A. Loeb, "Chaos in the test particle eccentric Kozai–Lidov mechanism," *The Astrophysical Journal*, vol. 791, no. 2, p. 86, 2014.

[3] E. Forest, "Geometric integration for particle accelerators," *Journal of Physics A: Mathematical and General*, vol. 39, no. 19, p. 5321, 2006.

[4] R. Salmon, "Hamiltonian fluid mechanics," *Annual Review of Fluid Mechanics*, vol. 20, no. 1, pp. 225–256, 1988.
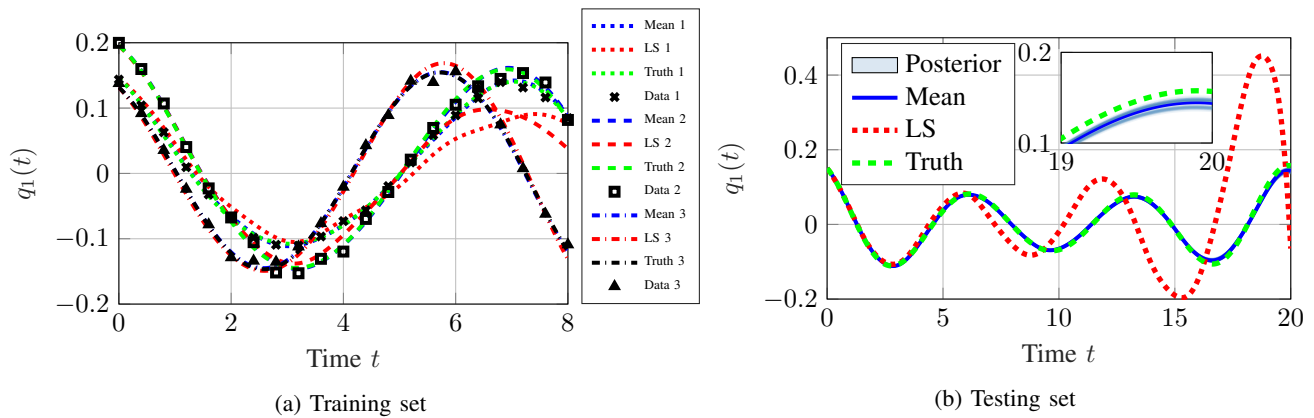
(a) Training set

(b) Testing set

Fig. 3: Left: The training data from 3 different ICs are shown alongside the point estimates from the LS [11] and Bayesian methods. The Bayesian method is more robust to the noisiness/sparsity of the data. Right: The posterior predictive distribution and its mean are compared to the LS point on a trajectory outside the training set. The Bayesian method gives a good prediction over $t = 20$, while the LS estimate deteriorates rapidly after about $t = 5$.

[5] J. Colliander, M. Keel, G. Staffilani, H. Takaoka, and T. Tao, "Transfer of energy to high frequencies in the cubic defocusing nonlinear Schrödinger equation," *Inventiones Mathematicae*, vol. 181, no. 1, pp. 39–113, 2010.

[6] S. Greydanus, M. Dzamba, and J. Yosinski, "Hamiltonian neural networks," *Advances in Neural Information Processing Systems*, vol. 32, pp. 15 379–15 389, 2019.

[7] Z. Chen, J. Zhang, M. Arjovsky, and L. Bottou, "Symplectic recurrent neural networks," in *International Conference on Learning Representations*, 2019.

[8] A. Zhu, P. Jin, and Y. Tang, "Deep Hamiltonian networks based on symplectic integrators," *arXiv preprint arXiv:2004.13830*, 2020.

[9] P. Jin, Z. Zhang, A. Zhu, Y. Tang, and G. E. Karniadakis, "SympNets: Intrinsic structure-preserving symplectic networks for identifying Hamiltonian systems," *Neural Networks*, vol. 132, pp. 166–179, 2020.

[10] H. K. Chu and M. Hayashibe, "Discovering interpretable dynamics by sparsity promotion on energy and the Lagrangian," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 2154–2160, 2020.

[11] K. Wu, T. Qin, and D. Xiu, "Structure-preserving method for reconstructing unknown Hamiltonian systems from trajectory data," *SIAM Journal on Scientific Computing*, vol. 42, no. 6, pp. A3704–A3729, 2020.

[12] T. Bertalan, F. Dietrich, I. Mezić, and I. G. Kevrekidis, "On learning Hamiltonian systems from data," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 29, no. 12, p. 121107, 2019.

[13] C. Offen and S. Ober-Blöbaum, "Symplectic integration of learned Hamiltonian systems," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 32, no. 1, p. 013122, 2022.

[14] N. Galioto and A. A. Gorodetsky, "Bayesian identification of Hamiltonian dynamics from symplectic data," in *2020 59th IEEE Conference on Decision and Control (CDC)*. IEEE, 2020, pp. 1190–1195.

[15] S. Xiong, Y. Tong, X. He, S. Yang, C. Yang, and B. Zhu, "Nonseparable symplectic neural networks," in *International Conference on Learning Representations*, 2020.

[16] R. Chen and M. Tao, "Data-driven prediction of general Hamiltonian dynamics via learning exactly-symplectic maps," in *International Conference on Machine Learning*. PMLR, 2021, pp. 1717–1727.

[17] N. Galioto and A. A. Gorodetsky, "Bayesian system ID: optimal management of parameter, model, and measurement uncertainty," *Nonlinear Dynamics*, vol. 102, no. 1, pp. 241–267, 2020.

[18] E. Hairer, C. Lubich, and G. Wanner, *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*. Springer Science & Business Media, 2006, vol. 31.

[19] H. Sharma, M. Patil, and C. Woolsey, "A review of structure-preserving numerical methods for engineering applications," *Computer Methods in Applied Mechanics and Engineering*, vol. 366, p. 113067, 2020.

[20] T. M. Cherry, "V. On periodic solutions of Hamiltonian systems

differential equations," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 227, no. 647-658, pp. 137–221, 1928.

[21] A. Casti, P. Morrison, and E. Spiegel, "Negative energy modes and gravitational instability of interpenetrating fluids," *Annals of the New York Academy of Sciences*, vol. 867, no. 1, pp. 93–108, 1998.

[22] P. Joarder, V. Nakariakov, and B. Roberts, "A manifestation of negative energy waves in the solar atmosphere," *Solar Physics*, vol. 176, no. 2, pp. 285–297, 1997.

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

[24] E. T. Jaynes, *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

[25] S. Särkkä, *Bayesian Filtering and Smoothing*. Cambridge University Press, 2013.

[26] S. Noh, "Posterior inference on parameters in a nonlinear DSGE model via Gaussian-based filters," *Computational Economics*, pp. 1–47, 2019.

[27] C. Drovandi, R. G. Everitt, A. Golightly, and D. Prangle, "Ensemble MCMC: accelerating pseudo-marginal MCMC for state space models using the ensemble Kalman filter," *Bayesian Analysis*, vol. 17, no. 1, pp. 223–260, 2022.

[28] M. Khalil, A. Sarkar, S. Adhikari, and D. Poirel, "The estimation of time-invariant parameters of noisy nonlinear oscillatory systems," *Journal of Sound and Vibration*, vol. 344, pp. 81 – 100, 2015.

[29] F. Yang, Z. Wang, and Y. Hung, "Robust Kalman filtering for discrete time-varying uncertain systems with multiplicative noises," *IEEE Transactions on Automatic Control*, vol. 47, no. 7, pp. 1179–1183, 2002.

[30] F. Gustafsson, "Particle filter theory and practice with positioning applications," *IEEE Aerospace and Electronic Systems Magazine*, vol. 25, no. 7, pp. 53–82, 2010.

[31] H. Haario, M. Laine, A. Mira, and E. Saksman, "DRAM: Efficient adaptive MCMC," *Statistics and Computing*, vol. 16, no. 4, pp. 339–354, 2006.

[32] Y. D. Zhong, B. Dey, and A. Chakraborty, "Symplectic ODE-net: Learning Hamiltonian dynamics with control," in *International Conference on Learning Representations*, 2019.

[33] H. Sharma, Z. Wang, and B. Kramer, "Hamiltonian operator inference: Physics-preserving learning of reduced-order models for canonical Hamiltonian systems," *Physica D: Nonlinear Phenomena*, vol. 431, p. 133122, 2022.