Research

Cite this article: Yang Y, Bhourri MA, Perdikaris P. 2020 Bayesian differential programming for robust systems identification under uncertainty. *Proc. R. Soc. A* **476**: 20200290.
<https://doi.org/10.1098/rspa.2020.0290>

Received: 17 April 2020

Accepted: 23 October 2020

Subject Areas:

applied mathematics, computational mathematics, computer modelling and simulation

Keywords:

machine learning, dynamical systems, uncertainty quantification, model discovery

Author for correspondence:

Paris Perdikaris

e-mail: pqp@seas.upenn.edu

Bayesian differential programming for robust systems identification under uncertainty

Yibo Yang, Mohamed Aziz Bhourri, Paris Perdikaris

Department of Mechanical Engineering and Applied Mechanics, University of Pennsylvania, Philadelphia, PA 19104, USA

PP, 0000-0002-2816-3229

This paper presents a machine learning framework for Bayesian systems identification from noisy, sparse and irregular observations of nonlinear dynamical systems. The proposed method takes advantage of recent developments in differentiable programming to propagate gradient information through ordinary differential equation solvers and perform Bayesian inference with respect to unknown model parameters using Hamiltonian Monte Carlo sampling. This allows an efficient inference of the posterior distributions over plausible models with quantified uncertainty, while the use of sparsity-promoting priors enables the discovery of interpretable and parsimonious representations for the underlying latent dynamics. A series of numerical studies is presented to demonstrate the effectiveness of the proposed methods, including nonlinear oscillators, predator–prey systems and examples from systems biology. Taken together, our findings put forth a flexible and robust workflow for data-driven model discovery under uncertainty. All codes and data accompanying this article are available at <https://bit.ly/34FOJMj>.

1. Introduction

In the era of big data, dynamical systems discovery has received a lot of attention, primarily because of the significant growth of accessible data across different scientific disciplines, including systems biology [1], bio-medical imaging [2], fluid dynamics [3], climate modelling [4] and physical chemistry [5]. Extracting a set of features that are interpretable and predictive is crucial for developing physical insight and gaining a better

understanding of the natural phenomena under study [6]. Perhaps more importantly, this can enable the reliable forecasting of future states and subsequently lead to effective intervention strategies for design and control of complex systems [7–9].

Machine learning methods and data-driven modelling techniques have already proven their utility in solving high-dimensional problems in computer vision [10], natural language processing [11], etc. Owing to their capability of extracting features from high-dimensional and multi-fidelity noisy data [12,13], these methods are also gaining attraction in modelling and simulating physical and biological systems. The evolution of such systems can be typically characterized by differential equations, and several techniques have been developed to construct predictive algorithms that can synergistically combine data and mechanistic prior knowledge. Such scientific machine learning approaches are currently employed to distil dynamics from time-series data [14–21], infer the solution of differential equations [22–27], infer parameters, latent variables and unknown constitutive laws [3,28–31], as well as tackle forward and inverse problems in complex application domains including cardiovascular flow dynamics, [32], metamaterials [33], cardiac electrophysiology [34], etc.

Specific to systems identification, most recent data-driven approaches [14–17,21] rely heavily on the quality of the observations and are not designed to return predictions with quantified uncertainty. For instance, the sparse regression methods put forth in [17,18] can exhibit unstable behaviour if the data are highly noisy and are not able to directly digest time-series data with irregular sampling frequency or missing values. On the other hand, recent approaches leveraging differentiable programming [14–16] can support irregularly sampled data, but are only designed to provide point estimates for the discovered dynamics with no characterization of predictive uncertainty. Such a lack of robustness and missing capabilities may limit the use of existing techniques to idealized settings and pose the need for a more flexible framework that can effectively accommodate noisy, sparse and irregularly sampled data to infer posterior distributions over plausible models, and subsequently yield robust future forecasts with quantified uncertainty.

In this work, we aim to address the aforementioned capability gaps by formulating a fully Bayesian framework for robust systems identification from imperfect time-series data. Our specific contributions can be summarized in the following points:

- We leverage recent developments in differentiable programming [14–16] to enable gradient back-propagation through numerical ordinary differential equation (ODE) solvers, and use this information to construct accelerated Hamiltonian Monte Carlo (HMC) schemes for Bayesian inference.
- The proposed workflow is computationally efficient, end-to-end differentiable and can directly accommodate sparse, noisy and irregularly sampled time-series data.
- Equipped with sparsity-promoting priors, we can recover interpretable and parsimonious representations for the latent dynamics, while entire posterior distributions over plausible models are obtained.
- This probabilistic formulation is key for safeguarding against erroneous data and incomplete model parametrizations, as well as for producing reliable future forecasts with quantified uncertainty.
- We demonstrate enhanced capabilities and robustness against state-of-the-art methods for systems identification [17,18] across a range of benchmark problems.

Taken together, our findings put forth a novel, flexible and robust workflow for data-driven model discovery under uncertainty that can potentially lead to improved algorithms for robust forecasting, control and model-based reinforcement learning of complex systems.

The rest of this paper is organized as follows. Section 2 presents the proposed method and its corresponding technical ingredients. In §2b, a review of Bayesian inference and HMC sampling is given. In §2a, we provide details on how the proposed method can effectively propagate gradient information through ODE solvers leveraging differential programming. In §2e, we discuss a

simple but essential step for pre-processing the observed data in order to obtain robust training behaviour. In §3, the effectiveness of the proposed method is tested on a series of numerical studies, including nonlinear oscillators, predator–prey systems and a realistic example in systems biology. Finally, in §4, we summarize our key findings, discuss the limitations of the proposed approach and carve out directions for future investigation.

2. Methods

This section provides a comprehensive overview of the key ingredients that define our work, namely differential programming via neural ordinary differential equations (neuralODEs) [16] and Bayesian inference with HMC sampling [35]. Our presentation is focused on describing how these techniques are interfaced to obtain a novel, efficient and robust workflow for parsimonious model discovery from imperfect time-series data.

(a) Differentiable programming

In their original work, Chen *et al.* [16] introduced a general framework for propagating gradient information through classical numerical solvers for ODEs that blends classical adjoint methods [36] with modern developments in automatic differentiation [37]. To illustrate the main concepts, consider a general dynamical system of the form

$$\frac{dx}{dt} = f(x, t; \theta), \quad (2.1)$$

where $x \in \mathbb{R}^D$ denotes the state space of the D -dimensional dynamical system and θ is a vector of unknown parameters that parametrizes the latent dynamics $f: \mathbb{R}^D \rightarrow \mathbb{R}^D$. A systems identification task is now summarized as follows. Given some observations x_i , $i = 1, \dots, n$ evaluated at time instances t_i , $i = 1, \dots, n$, one would like to learn the θ that best parametrizes the underlying dynamics. A typical approach for identifying these optimal parameters is to define a loss function g that measures the discrepancy between the observed data x_{i+1} and the model's predictions $\hat{x}_{i+1}$ for a given θ , i.e.

$$\mathcal{J}(\theta) = \sum_{i=1}^{n-1} g(x_{i+1}, h_{\theta}(x(t_i))), \quad (2.2)$$

where $h_{\theta}(x(t_i)) = \hat{x}_{i+1}$ is the predicted value under a given set of estimated model parameters θ obtained by integrating the dynamical system with some ODE solver. $g(\cdot)$ could be any metric to evaluate the distance/discrepancy between the true value and the predicted one (e.g. L_1 loss, L_2 loss [38] or KL divergence [39], Wasserstein distance [40]). A sufficient way to minimize the loss function is through gradient descent [41,42]; however, appropriate methods need to be employed for effectively computing the gradient of the loss function with respect to the parameters, namely $\partial L / \partial \theta$. This is done by defining the adjoint of the dynamical system as $a(t) = \partial L / \partial x(t)$. Then, the dynamical system describing the evolution of the adjoint can be derived as [16,36]

$$\frac{da(t)}{dt} = -a(t)^T \frac{\partial f(x(t), t, \theta)}{\partial x}. \quad (2.3)$$

Note that the adjoint $a(t)$ can be computed by an additional call to the chosen ODE solver, and the target derivative $\partial L / \partial \theta$ can then be computed as

$$\frac{\partial L}{\partial \theta} = - \int_{t_1}^{t_0} a(t)^T \frac{\partial f(x(t), t, \theta)}{\partial \theta} dt, \quad (2.4)$$

where $\partial f(x(t), t, \theta) / \partial \theta$ can be evaluated via automatic differentiation [16,37]. The main advantages of this approach can be summarized in the following points:

- The data do not need to be collected on a regular time grid.

- The time step Δt_i between an observed data pair $\{x(t_i), x(t_i + \Delta t_i)\}$ can be relatively large. Within each Δt_i , a classical numerical scheme can be used to integrate equation (2.4), where $\Delta t_i = N_i dt$ is discretized in N_i equal spaced steps, with the step size dt being typically chosen according to the stability properties of the underlying ODE solver.
- As this set-up only assumes dependency between individual input–output pairs $\{x(t_i), x(t_i + \Delta t_i)\}$, the observed time-series data do not need to be continuous and could be selected from different time intervals.

The specific choices of Δt_i , dt and N_i will be discussed for each of the different examples given in this work. In general, the choice of N is made based on the following trade-off between accuracy and computational complexity. To this end, small N may lead to a less accurate model, while large N would lead to a massive computational graph that can significantly slow down model training.

The unknown model parameters θ can be estimated by minimizing the appropriate loss function. Throughout this work, the following L_2 loss is employed:

$$\mathcal{J}(\theta) = \sum_{i=1}^n \|x(t_i + \Delta t_i) - h_{\theta}(x(t_i))\|^2, \quad (2.5)$$

where $h_{\theta}(x(t_i))$ denotes the output of a numerical ODE solver. Throughout this work, we use the classical fourth-order Runge–Kutta method [43], although more general choices can be employed [14]. The training dataset consists of pairs $\{(x(t_1), x(t_1 + \Delta t_1)), (x(t_2), x(t_2 + \Delta t_2)), \dots, (x(t_n), x(t_n + \Delta t_n))\}$. To simplify notation, let $X(t)$ be defined as $X(t) = \{x(t_1), x(t_2), \dots, x(t_n)\}$, such that $X(t + \Delta t) = \{x(t_1 + \Delta t_1), x(t_2 + \Delta t_2), \dots, x(t_n + \Delta t_n)\}$, then the training dataset is given by: $\mathcal{D} = \{X(t), X(t + \Delta t)\}$.

Note that the aforementioned workflow is only capable of producing deterministic point estimates for the unknown parameters θ , typically corresponding to a local minimum of equation (2.5). In many practical cases, it is desirable to obtain a distribution over plausible parameter configurations that can effectively characterize the uncertainty in the estimates due to noise and sparsity in the observed data, as well as potential misspecifications in the model parametrization. A framework for accounting for such uncertainties can be constructed by adopting a Bayesian approach via the use of effective sampling algorithms for approximating a posterior distribution over the unknown model parameters $p(\theta|\mathcal{D})$, as discussed in the next section.

(b) Bayesian inference with Hamiltonian Monte Carlo

The Bayesian formalism provides a natural way to account for uncertainty, while also enabling the injection of prior information for the unknown model parameters θ (e.g. sparsity), as well as for modelling sparse and noisy observations in the training dataset. Perhaps more importantly, it enables the complete statistical characterization for all inferred parameters in the model. The latter is encapsulated in the posterior distribution, which can be factorized as

$$p(\gamma, \lambda, \theta | X(t + \Delta t), X(t)) \propto p(X(t + \Delta t) | X(t), \theta, \gamma) p(\theta | \lambda) p(\gamma) p(\lambda), \quad (2.6)$$

where $p(X(t + \Delta t) | X(t), \theta, \gamma)$ is a likelihood function that measures the discrepancy between the observed data and the model's predictions for a given set of parameters θ , $p(\theta | \lambda)$ is a prior distribution parametrized by a set of parameters λ that can help encode any domain knowledge about the unknown model parameters θ and γ contains a set of parameters that aim to

characterize the noise process that may be corrupting the observations. In this work, a hierarchical Bayesian approach corresponding to the following likelihood and priors is employed:

$$\left. \begin{aligned} p(X(t + \Delta t)|X(t), \theta, \gamma) &= \prod_{i=1}^N \mathcal{N}(x(t_i + \Delta t_i) | f(x(t_i), t_i, \theta), \gamma^{-1}), \\ p(\theta|\lambda) &= \text{Laplace}(\theta|0, \lambda^{-1}), \\ p(\log \lambda) &= \text{Gam}(\log \lambda | \alpha_1, \beta_1) \\ p(\log \gamma) &= \text{Gam}(\log \gamma | \alpha_2, \beta_2). \end{aligned} \right\} \quad (2.7)$$

and

The use of a Gaussian likelihood stems from assuming a simple uncorrelated Gaussian noise model with zero mean and precision γ . The Gamma distribution is a common choice for the prior distributions over the unknown precision parameters λ and γ . For additional motivation and alternative choices, the interested reader is referred to [44–48]. Moreover, the logarithm of the precision variables λ and γ is used to ensure that their estimated values remain positive during model training. Finally, the Laplace prior over the unknown model parameters θ [49] can promote sparsity in the inferred model representations and enable an effective reduction in the influence of any irrelevant parameters—a key feature for recovering interpretable and parsimonious representations [17]. In this work, ‘interpretability’ refers to the ability to relate the identified terms of the dynamics to their physical meaning and effects. This is particularly interesting for physical systems whose dynamics can be described with different terms that can have numerous orders of magnitude and hence should be included in the model, while the proposed Bayesian differential programming (BDP) method should be able to identify the most relevant/important ones [17]. The latter are then interpreted based on their physical meaning.

The posterior distribution defined in equation (2.7) is not analytically tractable in general owing to the modelling assumptions on the likelihood and priors, as well as to the presence of nonlinearity in the latent dynamics of equation (2.1). Typically, sampling from this unnormalized distribution is difficult and computationally expensive, especially when the dimension of θ is large. HMC [35] is a powerful tool to handle Bayesian inference tasks in high dimensions by using gradient information to effectively generate approximate posterior samples. To illustrate the key ideas behind HMC sampling, let us denote $\Theta = [\theta, \log(\lambda), \log(\gamma)]$ as the vector including all the unknown parameters that need to be inferred from data. The starting point for building an HMC sampler is to define a Hamiltonian function $H = U(\Theta) + V(v)$, where $U(\Theta)$ is the potential energy of the original system, which is usually taken as the logarithm of the unnormalized distribution in equation (2.7), and $V(v)$ is the kinetic energy of the system introduced by the auxiliary velocity variables $v := d\Theta/dt$. The evolution of Θ and v can be expressed by taking gradients of the Hamiltonian as

$$\frac{dv}{dt} = -\frac{\partial H}{\partial \Theta}, \quad \frac{d\Theta}{dt} = \frac{\partial H}{\partial v}. \quad (2.8)$$

A Markov chain that converges to the stationary distribution of equation (2.6) can be simulated by integrating this dynamical system using an energy-preserving leapfrog scheme [35]. Note here that computing the gradient $\partial H(t)/\partial \Theta$ is not trivial as the evaluation of the likelihood function involves the call to an ODE solver. This difficulty can be directly addressed via the differentiable programming approach discussed in §2a to effectively enable gradient back-propagation through the ODE solver. Note that all parameters in Θ can be updated either simultaneously or separately for θ and $\{\lambda, \gamma\}$ using a Metropolis-within-Gibbs scheme [50,51].

(c) Learning dynamics with Bayesian differential programming

Here, we distinguish between three different problem settings that cover a broad range of practical applications. The first class consists of problems in which the model form of the underlying latent dynamics is completely unknown. In this case, one can parametrize the unknown dynamical system using black-box function approximators such as deep neural

networks [15,16,20], or aim to distil a more parsimonious and interpretable representation by constructing a comprehensive dictionary over all possible interactions and try to infer a predictive, yet minimal, model form [17,18,21]. The second class of problems contains cases where a model form for the underlying dynamics is prescribed by domain knowledge, but a number of unknown parameters need to be calibrated in order to accurately explain the observed data [31,52]. Finally, a third class of problems arises as a hybrid of the aforementioned settings, in which some parts of the model form may be known from domain knowledge, but there exists additional functional terms that need to be inferred [14]. As detailed in the following sections, the proposed workflow can seamlessly accommodate all of the aforementioned cases in a unified fashion, while remaining robust with respect to incomplete model parametrizations as well as imperfections in the observed data. To this end, a general framework can be constructed by parametrizing the unknown dynamics as

$$\frac{dx}{dt} = \underbrace{A\varphi(x)}_{\text{dictionary}} + \underbrace{f_w(x)}_{\text{black box}}, \quad (2.9)$$

where $A \in \mathbb{R}^{D \times K}$ represents a matrix of $D \times K$ unknown coefficients, with K being the length of a dictionary $\varphi(x) \in \mathbb{R}^K$, that may or may not be constructed using domain knowledge. Specifically, $\varphi(x)$ represents the possible terms that may appear on the right-hand side of the ODE, which could encapsulate a known model form or, more generally, a prescribed dictionary of features (e.g. polynomials, Fourier modes, etc. and combinations thereof) [17,18,53]. On the other hand, $f_w(x)$ denotes a black-box function approximator (e.g. a neural network) parametrized by w that aims to account for any missing interactions that are not explicitly captured by $\varphi(x)$ [14]. It is true that interpretability is lost to some extent when a deep learning block (e.g. a neural network) is introduced into the identification pipeline. Yet, black-box parts of the model only serve as a ‘closure’ that aims to capture parts of the physics that are *a priori* unknown and cannot be captured by a dictionary representation (although they can have an important contribution in the systems’ dynamics).

On the other hand, domain knowledge can play a crucial role in enhancing the efficiency of the resulting inference scheme as it can effectively reduce the size of the dictionary and, consequently, the amount of data required to train the model [23,24]. Such knowledge is also critical to constrain the space of admissible solutions such that key physical principles are faithfully captured by the inferred model (e.g. convergence to equilibrium limit cycles in chemical systems [54], conservation of mass and momentum in fluid dynamics [30]).

Although the use of dictionary learning (with or without specific domain knowledge) offers a flexible paradigm for recovering interpretable dynamic representations, it may not always be sufficient to explain the observed data, as important terms may be missing from the model parametrization. To address this shortcoming one can try to capture these missing interactions via the use of closure terms that often lack physical intuition, and hence can be represented by a black-box function approximator $f_w(x)$ with parameters w , such as a deep neural network [14]. Under this set-up, the algorithmic framework outlined in §2a,b is employed to jointly perform probabilistic inference over plausible sets of model parameters $\theta := \{A, w\}$ that yield interpretable, parsimonious and predictive representations, as well as the precision parameters λ and γ of the Bayesian hierarchical model in equation (2.7).

(d) Generating forecasts with quantified uncertainty

The goal of the HMC algorithm described in §2b is to produce a faithful set of samples that concentrate in regions of high probability in the posterior distribution $p(\theta, \lambda, \gamma | \mathcal{D})$. Approximating this distribution is central to the proposed workflow as it enables the generation of future forecasts $x^*(t)$ with quantified uncertainty via computing the predictive posterior distribution

$$p(x^*(t) | \mathcal{D}, x_0, t) = \int p(x^*(t) | \theta, x_0, t) p(\theta | \mathcal{D}) d\theta. \quad (2.10)$$

This predictive distribution provides a complete statistical characterization for the forecasted states by encapsulating epistemic uncertainty in the inferred dynamics, as well as accounting for the fact that the model was trained on a finite set of noisy observations. This allows the generation of plausible realizations of $x^*(t)$ by sampling from the predictive posterior distribution as

$$x^*(t) = h_{\theta}(x_0, t) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \gamma^{-1}), \quad \theta, \lambda, \gamma \sim p(\theta, \lambda, \gamma | \mathcal{D}), \quad (2.11)$$

where θ , λ and γ are approximate samples from $p(\theta, \lambda, \gamma | \mathcal{D})$ computed during model training via HMC sampling, $h_{\theta}(x_0, t)$ denotes any numerical integrator that takes some initial condition x_0 and predicts the system's state at any time t , and ϵ accounts for the noise corrupting the observations used during model training. Moreover, the maximum *a posteriori* (MAP) estimate of the model parameters is given as follows:

$$\theta_{\text{MAP}}, \lambda_{\text{MAP}}, \gamma_{\text{MAP}} = \arg \max_{\theta, \lambda, \gamma} p(\theta, \lambda, \gamma | \mathcal{D}). \quad (2.12)$$

This is used to obtain a point-estimate prediction of the predicted states $\hat{x}_{\text{MAP}}(t)$ defined as

$$x_{\text{MAP}}^*(t) = h_{\theta_{\text{MAP}}}(x_0, t). \quad (2.13)$$

Finally, it is straightforward to use the posterior samples of $\theta \sim p(\theta | \mathcal{D})$ to approximate the first- and second-order statistics of the predicted states $x^*(t)$ for any given initial condition x_0 as

$$\hat{\mu}_{x^*}(x_0, t) = \int h_{\theta}(x_0, t) p(\theta | \mathcal{D}) d\theta \approx \frac{1}{N_s} \sum_{i=1}^{N_s} h_{\theta_i}(x_0, t) \quad (2.14)$$

and

$$\hat{\sigma}_{x^*}^2(x_0, t) = \int [h_{\theta}(x_0, t) - \hat{\mu}_{x^*}(x_0, t)]^2 p(\theta | \mathcal{D}) d\theta \approx \frac{1}{N_s} \sum_{i=1}^{N_s} [h_{\theta_i}(x_0, t) - \hat{\mu}_{x^*}(x_0, t)]^2, \quad (2.15)$$

where N_s denotes the number of samples drawn from the Hamiltonian Markov chain used to simulate the posterior, i.e. $\theta_i \sim p(\theta | \mathcal{D})$, $i = 1, \dots, N_s$. Note that higher-order moments are also readily computable in a similar manner.

(e) Model initialization and data pre-processing

To promote robustness and stability in the training of the proposed machine learning pipeline, users should be cognizant of several important aspects. First, although the proposed Bayesian approach can naturally safeguard against over-fitting, it is important that a reasonable amount of training data are provided—relative to the complexity of the system—in order to mitigate any effects of prior misspecification. Second, the training data should be appropriately normalized in order to prevent gradient pathologies during back-propagation [55]. The specific utility of this step will be demonstrated in the numerical examples presented in this work. This step is carried out using a standard normalization of the form

$$\tilde{x} = \frac{x}{\sigma_x}, \quad (2.16)$$

where σ_x is the dimension-wise standard deviation of the training data and the division is an element-wise operation. Note that this modification directly implies that the assumed parametrization of the underlying dynamical system also needs to be normalized accordingly (see §3 for a more detailed discussion). A third important remark here is that the noise precision γ obtained from the model aims to reflect the noise level in the observed data. However, it may not be the true noise level because the initial condition of the ODE can also be noisy. This point will be discussed further in §3.

Another important point relates to the initialization of the HMC Markov chain sampler. To this end, in order to mitigate poor mixing and convergence to local minima, a preconditioning step is considered to find a reasonable initial guess for the unknown variables θ that parametrize the

underlying latent dynamics. This step is typically carried out by minimizing the reconstruction loss of the training data using an L_1 regularization,

$$\mathcal{L}(\theta) = \frac{1}{n} \sum_{i=1}^n \|x_i - \hat{x}_i\|^2 + \beta \|\theta\|_1, \quad (2.17)$$

using a small number ($\mathcal{O}(10^3)$) of stochastic gradient descent iterations. Note that this essentially aims at obtaining a rough point estimate for θ , where the use of L_1 regularization stems from employing a sparsity-promoting Laplace prior. This preconditioning step is closely related to the SINDy algorithm of Brunton *et al.* [17], but without the limitations of requiring training data with small noise amplitude and sampled on regular time grid with small a time step. Moreover, the numerical experiments carried out indicate that tuning the hyper-parameter β has almost no effect on the obtained results for all the problems considered in this work, as this simply serves as an initialization step for the HMC sampler. Hence, in all examples considered, β is taken equal to 1. The minimization of equation (2.17) is carried out using stochastic Adam updates [56] with a learning rate of $\mathcal{O}(10^{-2})$.

Note that this preconditioning is precisely equivalent to the MAP estimation for all model parameters θ excluding the noise precision parameters λ and γ , which are initialized as detailed below. Specifically, for all the problems considered in §3, the parameters α_i 's and β_i 's of the prior Gamma distributions are chosen to be 1. Moreover, the precision of the Gaussian noise distribution γ is initialized as follows. If the preconditioning precision is larger than $\exp(6)$, which means that the training data appear to be nearly noise-free, the initial guess for γ is set to $\exp(6)$ to avoid numerical stagnancy of the HMC sampler. Otherwise, if the preconditioning precision is less than $\exp(6)$, it is used as the initial guess for γ . This empirical initialization strategy has a positive effect in accelerating the convergence of the HMC sampler for all the examples considered in §3. In all of the numerical examples, the HMC step size is taken as $\epsilon = 10^{-4}$, while the number of leapfrog steps to integrate the Hamiltonian dynamics in equation (2.8) is fixed to $L = 10$. While this choice of ϵ is the safest choice, one can increase its value as long as not too many samples are rejected during model training. Alternatively, more sophisticated HMC samplers that allow for adaptively tuning the step size can be employed [57]. Finally, in all examples, the Markov chains are simulated for 5000 steps, while the last $N_s = 2000$ samples produced by HMC are used to compute the response statistics (see equation (2.14)). We also note that the computational complexity is approximately constant for all examples considered regardless of the problem's dimension, and takes around 3 h on a four-core MacBook Pro laptop (with a 3.5 GHz Intel CPU and 16 GB of RAM).

3. Results

In this section, a comprehensive collection of numerical studies that aim to illustrate the key contributions of this work is presented and placed in the context of the existing SINDy framework of Brunton *et al.* [17], which is currently considered as a state-of-the-art method for dictionary learning of dynamical systems. Specifically, we expand on four benchmark problems that cover all possible cases discussed in §2c in terms of parametrizing the latent dynamics using a dictionary, domain knowledge, black-box approximations, or a combination thereof. The algorithmic settings used across all cases follow the discussion provided in §2e, unless otherwise noted. All code and data presented in this section will be made publicly available at <https://github.com/PredictiveIntelligenceLab/BayesianDifferentiableProgramming>.

(a) Dictionary learning for a two-dimensional nonlinear oscillator

Let us start with a pedagogical example on dictionary learning for inferring the dynamics of a two-dimensional damped oscillator from scattered time-series data [20]. The exact system

Figure 1. Two-dimensional damped oscillator with low-resolution training data. (a) Learned dynamics versus the true dynamics and the training data for $x_1(t)$. (b) SINDy's prediction for $x_1(t)$ versus the true dynamics and the training data. (c) Learned dynamics versus the true dynamics and the training data for $x_2(t)$. (d) SINDy's prediction for $x_2(t)$ versus the true dynamics and the training data. (Online version in colour.)

dynamics are given by

$$\left. \begin{aligned} \frac{dx_1}{dt} &= \alpha x_1^3 + \beta x_2^3 \\ \frac{dx_2}{dt} &= \gamma x_1^3 + \delta x_2^3 \end{aligned} \right\} \quad (3.1)$$

and

The goal here is to recover this dynamical system directly from data using a dictionary parametrization containing polynomial terms with up to third-order interactions, taking the form

$$\begin{bmatrix} \frac{dx_1}{dt} \\ \frac{dx_2}{dt} \end{bmatrix} = A\varphi(x) = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} & a_{16} & \mathbf{a_{17}} & a_{18} & a_{19} & \mathbf{a_{110}} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} & a_{26} & \mathbf{a_{27}} & a_{28} & a_{29} & \mathbf{a_{210}} \end{bmatrix} \varphi(x), \quad (3.2)$$

where $\varphi(x) = [1, x_1, x_2, x_1^2, x_1x_2, x_2^2, x_1^3, x_1^2x_2, x_1x_2^2, x_2^3]^T$ and the a_{ij} 's are unknown scalar coefficients that will be estimated using the proposed BDP method. The model's active non-zero parameters are highlighted with red colour for clarity. The goal is to infer a posterior distribution for the parameters $\theta := \{A\}$, while it is obvious that the coefficient matrix A and dictionary $\varphi(x)$ will increase in size with the order of the polynomial features used to parametrize the system. In all presented experiments, a set of training data is generated by simulating the exact dynamics of equation (3.1) in the time interval $t \in [0, 20]$ with the initial condition $(x_1, x_2) = (2, 0)$ and with $\alpha = -0.1, \beta = 2.0, \gamma = -2.0$ and $\delta = -0.1$. In what follows, the performance of the proposed algorithms is investigated with respect to the temporal resolution and the noise level in the observed data. Moreover, the analysis is provided with a comprehensive comparison with the SINDy algorithm of Brunton *et al.* [17].

(i) Effects of sparsity in the training data

To examine the performance of the proposed methods with respect to sparsity in the training data, a dataset is generated by integrating the exact dynamics with a relatively large $dt = 0.067$,

Figure 2. Two-dimensional damped oscillator with low-resolution training data. (a) Phase plot of the training data and the true trajectory. (b) Posterior distribution of the inferred active model parameters. (Online version in colour.)

Table 1. Dictionary learning for a two-dimensional nonlinear oscillator: MAP estimation of the inferred model parameters using low-resolution training data ($dt = 0.067$).

a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a_{16}	a_{17}	a_{18}	a_{19}	a_{110}
−0.0052	0.0057	−0.0095	0.0094	0.0116	0.0057	−0.097	−0.0093	−0.054	2.092
a_{21}	a_{22}	a_{23}	a_{24}	a_{25}	a_{26}	a_{27}	a_{28}	a_{29}	a_{210}
0.0048	0.0195	−0.0072	−0.021	0.0024	−0.0075	−2.08	0.0280	−0.014	−0.099

Table 2. Dictionary learning for a two-dimensional nonlinear oscillator: point estimates for the dictionary coefficients obtained by the SINDy algorithm [17] using low-resolution training data ($dt = 0.067$).

a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a_{16}	a_{17}	a_{18}	a_{19}	a_{110}
0	0	0	0	0	0	0	0	0	2.10
a_{21}	a_{22}	a_{23}	a_{24}	a_{25}	a_{26}	a_{27}	a_{28}	a_{29}	a_{210}
0	0	0	0	0	0	−1.87	0	0	0

such that there are only $n = 300$ training data pairs. In order to establish a comparison against the SINDy algorithm [17], the training data are assumed noise-free and sampled on a regular temporal grid, as required by the SINDy set-up [17].

As seen in figures 1 and 2, the SINDy algorithm fails to identify the underlying system, while the proposed approach remains robust owing to the Bayesian formulation outlined in §2b and enables a faithful statistical characterization of the latent system dynamics even under sparse observations. Indeed, the proposed method not only has the capability to accurately identify the parameters even for relatively large dt (table 1), but also provides reasonable uncertainty estimates for the extrapolated long-term forecasts. By contrast, the SINDy algorithm gives inaccurate estimations for the dictionary parameters (table 2), consequently leading to large errors in the forecasted system states. As shown in figure 1*b,d*, the estimated trajectories clearly deviate from the exact solution since its oscillatory frequency is considerably higher than the exact one, and it does not capture the decay in the oscillations' peaks. Moreover, SINDy's results highly depend on the choice of the sequential least-squares threshold parameter value, and slightly different values of this hyper-parameter may give significantly different results.

Figure 3. Two-dimensional damped oscillator with noisy training data. (a) Learned dynamics versus the true dynamics and the training data for $x_1(t)$. (b) SINDy's prediction for $x_1(t)$ versus the true dynamics and the training data. (c) Learned dynamics versus the true dynamics and the training data for $x_2(t)$. (d) SINDy's prediction for $x_2(t)$ versus the true dynamics and the training data. (Online version in colour.)

Figure 4. Two-dimensional damped oscillator with noisy training data. (a) Phase plot of the training data and the true trajectory. (b) Posterior distribution of the inferred active model parameters. (Online version in colour.)

In order to further investigate how SINDy and the proposed approach perform at different temporal resolutions, we present the following systematic study. We consider four different time steps to compare the proposed BDP method's identification efficiency with SINDy by taking: $dt = 0.1$, $dt = 0.067$, $dt = 0.04$, $dt = 0.02$ and $dt = 0.005$. The obtained results are shown in table 3. The proposed method is able to accurately identify all the parameters, while SINDy fails to do so for the largest time steps considered: $dt = 0.067$ and $dt = 0.1$, which correspond to a relatively sparse set of temporal observations (300 and 200 training points, respectively).

Table 3. Relative error of the parameters' identification for the spiral dynamics problem obtained with the proposed BDP method and with SINDy: the comparison is carried out using the MAP estimate of the proposed BDP method versus SINDy's output at different time step sizes: $dt = 0.1$, $dt = 0.067$, $dt = 0.04$, $dt = 0.02$, $dt = 0.005$.

dt no. training data	1.0×10^{-1} 200	6.7×10^{-2} 300	4.0×10^{-2} 500	2.0×10^{-2} 1000	5.0×10^{-3} 4000
SINDy a_{17}	1.0	1.0	4.6×10^{-3}	9.1×10^{-3}	2.6×10^{-4}
BDP a_{17}	2.6×10^{-2}	2.1×10^{-2}	5.0×10^{-2}	4.0×10^{-2}	2.0×10^{-2}
SINDy a_{27}	1.3×10^{-1}	1.4×10^{-2}	4.2×10^{-3}	9.6×10^{-4}	2.5×10^{-4}
BDP a_{27}	7.5×10^{-2}	4.4×10^{-2}	4.2×10^{-2}	2.8×10^{-2}	5.0×10^{-4}
SINDy a_{210}	1.0×10^{-2}	1.3×10^{-3}	9.8×10^{-4}	9.4×10^{-4}	2.5×10^{-4}
BDP a_{210}	8.7×10^{-2}	5.5×10^{-2}	2.5×10^{-2}	1.1×10^{-2}	1.5×10^{-3}
SINDy a_{110}	1.0	1.0	2.7×10^{-3}	2.2×10^{-4}	2.4×10^{-4}
BDP a_{110}	5.0×10^{-2}	1.4×10^{-2}	2.5×10^{-2}	2.7×10^{-2}	2.6×10^{-2}

Table 4. Dictionary learning for a two-dimensional nonlinear oscillator: MAP estimation of the inferred model parameters using noisy training data.

a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a_{16}	a_{17}	a_{18}	a_{19}	a_{110}
0.0042	-0.0059	-0.0045	-0.0072	-0.0078	-0.0071	-0.10	0.047	0.014	2.0
a_{21}	a_{22}	a_{23}	a_{24}	a_{25}	a_{26}	a_{27}	a_{28}	a_{29}	a_{210}
-0.0029	-0.010	-0.0073	0.0017	0.017	0.0064	-2.0	0.015	-0.049	-0.108

Table 5. Dictionary learning for a two-dimensional nonlinear oscillator: estimation of the SINDy's parameters using noisy training data.

a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a_{16}	a_{17}	a_{18}	a_{19}	a_{110}
0	0	0	0	0	0	-0.139	0	0	2.0
a_{21}	a_{22}	a_{23}	a_{24}	a_{25}	a_{26}	a_{27}	a_{28}	a_{29}	a_{210}
0	0	0	0	0	0	-1.97	0	-0.188	0

(ii) Effects of noise in the training data

In this section, the sensitivity of the proposed methods with respect to the presence of noise in the training data is investigated. To this end, a training dataset is generated with $n = 1000$ equi-spaced data pairs in $t \in [0, 20]$ by simulating the exact dynamics of equation (3.1) with $dt = 0.02$, and the observations are deliberately corrupted with uncorrelated Gaussian noise of the form $\mathcal{N}(0, 0.02^2)$ (figure 4a).

Figure 3 summarizes the predictions of the proposed Bayesian framework in comparison with the SINDy algorithm of Brunton *et al.* [17]. Moreover, the inferred dictionary parameters are provided for both methods in tables 4 and 5, respectively. Note that although the predicted trajectories of SINDy are quite close to the true trajectories, the identified parameters are quite different from the true model form. Even though SINDy employs a total variation diminishing (TVD) regularization to safeguard against small noise corruptions in the training data [17], it is still prone to providing inaccurate results in cases where the training data are imperfect. This limitation is addressed in the proposed framework by explicitly accounting for the effects of noise in the training data using the hierarchical Bayesian model in equation (2.6). This source

Figure 5. HMC convergence diagnostics for the two-dimensional damped oscillator parameters a_{17} , a_{27} , a_{110} , a_{210} . (a) The Geweke diagnostic test based on the chain of $N_t = 500$ samples: we compare the mean of the first 10% of the chain of N_t samples with the mean of 20 subsamples from the last 50% of the chain. (b) Auto-correlation as a function of the lag. (Online version in colour.)

of uncertainty is also effectively propagated through the system's dynamics to yield a sensible characterization of uncertainty in the predicted forecasts (figure 3a), via the inferred posterior distribution over the model parameters (figure 4b). Moreover, the resulting MAP estimates for the model parameters exhibit excellent agreement with the ground truth, as reported in table 4. This result illustrates the robust performance of the proposed framework in identifying interpretable and parsimonious system representations, even in the presence of noise in the training data.

(iii) Diagnostics of HMC convergence

To verify the statistical convergence of the HMC chains, we have employed the Geweke diagnostic and Gelman–Rubin tests [44,45]. To this end, here we provide a representative result for a time step size $dt = 0.005$. The Gelman–Rubin tests have provided values for the non-zero parameters (a_{17} : 1.016, a_{210} : 1.006, a_{110} : 1.001, a_{27} : 1.005) that are all very close to the optimal value of 1, as expected. In order to perform the Geweke diagnostic, we have considered a chain $N = 5000$ draws, using the first $N_b = 3000$ draws as a burn-in phase. We have also ‘thinned’ the remaining draws by randomly sampling one data point in every four consecutive steps of the chain. In order to implement the Geweke diagnostic, we considered the first 10% of the chain of $N_t = 500$ samples and compared their mean with the mean of 20 subsamples from the last 50% of the chain. As shown in figure 5a, the Geweke z-scores for the four non-zero parameters are well between -2 and 2 . The results of auto-correlation are given in figure 5b. The Geweke diagnostic, the Gelman–Rubin tests results and the auto-correlation values demonstrate a ‘healthy’ mixing behaviour of the HMC sampler.

(b) Parameter inference in a predator–prey system with irregularly sampled observations

This case study is designed to illustrate the capability of the proposed framework to accommodate noisy and irregularly sampled time-series data, a common practical setting that cannot be effectively addressed by SINDy and other popular data-driven systems identification methods [17,18,20,21,58]. To this end, a classical prey–predator system described by the Lotka–Volterra equations is considered

$$\left. \begin{aligned} \frac{dx_1}{dt} &= \alpha x_1 - \beta x_1 x_2 \\ \frac{dx_2}{dt} &= \delta x_1 x_2 - \gamma x_2, \end{aligned} \right\} \quad (3.3)$$

and

Figure 6. Parameter inference in a predator–prey system from irregularly sampled, noisy data. (a) Learned dynamics versus the true dynamics and the training data for x_1 . (b) Learned dynamics versus the true dynamics and the training data for x_2 . (Online version in colour.)

which is known to exhibit a stable limit cycle behaviour for $\alpha = 1.0$, $\beta = -0.1$, $\gamma = -1.5$ and $\delta = 0.75$. Without loss of generality, a dictionary that precisely contains the active terms of the system is considered. However, the $n = 1000$ training data pairs will be irregularly sampled in the interval $t \in [0, 25]$ by randomly subsampling an exact trajectory of system (3.3) starting from an initial condition set to $(x_1, x_2) = (5, 5)$. Moreover, the training data are perturbed by 3% white noise proportional to their standard deviation. Given this irregular and noisy training data, the goal is to demonstrate the performance of the proposed Bayesian framework in identifying the unknown model parameters $\theta := \{\alpha, \beta, \gamma, \delta\}$ with quantified uncertainty, as well as in producing sensible forecasts of extrapolated future states.

The results of this experiment are summarized in figures 6 and 7. It is evident that the proposed BDP approach (i) is able to provide an accurate estimation for the unknown model parameters, (ii) yields a MAP estimator with a predicted trajectory that closely matches the exact system's dynamics, (iii) returns a posterior distribution over plausible models that captures both the epistemic uncertainty of the assumed parametrization and the uncertainty induced by training on a finite amount of noisy training data, and (iv) propagates this uncertainty through the system's dynamics to characterize variability in the predicted future states.

Another interesting observation here is that the HMC sampler is very efficient in identifying the importance/sensitivity of each inferred parameter in the model. For instance, less important parameters have the highest uncertainty, as observed in the posterior density plots shown in figure 7b. Specifically, note how the posterior distribution of α and γ has a considerably larger standard deviation than the other parameters, implying that the evolution of this dynamical system is less sensitive with respect to these parameters.

(c) Safeguarding against model inadequacy: a damped pendulum case study

The purpose of this example is to demonstrate the effects of a misspecified model parametrization, and to highlight how the proposed Bayesian framework can help detect such problematic cases and safeguard against them. Such cases may arise when domain knowledge is insufficient to guide the selection of a parsimonious model form, as well as when important interaction terms may be missing from a dictionary representation. To illustrate the main ideas, a simple damped pendulum system described by the following equation is considered:

$$\left. \begin{aligned} \frac{dx_1}{dt} &= \gamma x_2 \\ \frac{dx_2}{dt} &= -\alpha x_2 - \beta \sin(x_1), \end{aligned} \right\} \quad (3.4)$$

and

Figure 7. Parameter inference in a predator–prey system from irregularly sampled, noisy data. (a) Phase plot of the training data and the true trajectory. (b) Posterior distribution of the inferred model parameters. (Online version in colour.)

with $\gamma = 1$, $\alpha = 0.2$ and $\beta = 8.91$. A set of sparse and irregularly sampled training data pairs can be generated by randomly subsampling a simulated trajectory of the exact dynamics in $t \in [0, 20]$, starting from an initial condition $(x_1, x_2) = (-1.193, -3.876)$. This imperfect dataset can be used to recover an interpretable model representation via dictionary learning, albeit here we deliberately choose to use an incomplete dictionary containing polynomial terms only up to first order, i.e.

$$\begin{bmatrix} \frac{dx_1}{dt} \\ \frac{dx_2}{dt} \end{bmatrix} = A\varphi(x) = \begin{bmatrix} a_{11} & a_{12} & \mathbf{a_{13}} \\ a_{21} & \underline{a_{22}} & \mathbf{a_{23}} \end{bmatrix} \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix}, \quad (3.5)$$

where the unknown model parameters are $\theta := \{a_{11}, a_{12}, a_{13}, a_{21}, a_{22}, a_{23}\}$, with the active terms being indicated with bold for clarity. Moreover, underlining is used to highlight a mismatched term in the incomplete dictionary, hinting at its erroneous capacity to approximate the true $\sin(x_1)$ term using just x_1 as a feature. It is evident that such a dictionary choice cannot faithfully capture the exact physics of the problem, and is hence destined to yield inaccurate predictions. Nevertheless, here we argue that this discrepancy between the true form and the assumed incomplete parametrization of the dynamical system can be effectively detected by the proposed Bayesian workflow via inspecting the inferred posterior distribution over all parameters $p(\theta, \lambda, \gamma | \mathcal{D})$, which is expected to exhibit high entropy in the presence of a misspecified model parametrization and imperfect training data.

The results of this experiment are summarized in figures 8 and 9. In particular, figure 8 shows the scattered training data, the exact simulated trajectory, the predicted trajectory corresponding to the MAP estimate of the inferred model parameters as well as a two-standard-deviation band around a set of plausible forecasted states. As expected, the use of a misspecified dictionary leads to an inferred model that has difficulty in fitting the observed data and yields predictions with high variability. This model inadequacy is also clearly captured in the posterior distribution over the inferred parameters shown in figure 9. Indeed, the inferred density for the β parameter exhibits very high variance, as the β coefficient crucially corresponds to the misspecified term in the dictionary. This is a direct indication that the inferred model is inadequate to capture the observed reality, leading to forecasts with inevitably large error bars. This innate ability of the proposed framework to detect model misspecification via a rigorous quantification of predictive uncertainty can prove crucial in risk-sensitive applications, and provides an important capability

Figure 8. Damped pendulum with irregularly sampled data. (a) Learned dynamics versus the true dynamics and the training data for $x_1(t)$. (b) Learned dynamics versus the true dynamics and the training data for $x_2(t)$. (Online version in colour.)

Figure 9. Damped pendulum with irregularly sampled data. (a) Phase plot of the training data and the true trajectory. (b) Posterior distribution of the model's parameters. (Online version in colour.)

Figure 10. Safeguarding against model inadequacy in a damped pendulum representation. (a) Learned dynamics versus the true dynamics and the training data for $x_1(t)$. (b) Learned dynamics versus the true dynamics and the training data for $x_2(t)$. A deep neural network is used to approximate any interactions missing from an incomplete dictionary representation. (Online version in colour.)

that is currently missing from the recently active literature on data-driven model discovery [14,17,18,20,21,53].

A natural question to now ask is: can we still recover an accurate predictive model even if the true dynamics can only be partially captured by the assumed dictionary representation? To address this question, we turn our attention to the hybrid learning setting discussed in §2c,

Figure 11. Safeguarding against model inadequacy in a damped pendulum representation. (a) Deep neural network (NN) approximation to the missing dictionary term $-\beta \sin(x_1)$. The shaded region indicates the range of data seen during model training. (b) Posterior distribution of the inferred active model parameters. (Online version in colour.)

in which some parts of the model can be captured by sparsely selecting interpretable terms from a dictionary, while other missing parts or closure terms can be accounted for via a black-box function approximator. To this end, the damped pendulum case study is revisited, and the dictionary learning parametrization is endowed with the ability to approximate the missing $\sin(x_1)$ term via a deep neural network as

$$\begin{bmatrix} \frac{dx_1}{dt} \\ \frac{dx_2}{dt} \end{bmatrix} = A\varphi(x) + f_w(x) = \begin{bmatrix} a_{11} + a_{12}x_1 + a_{13}x_2 + a_{14}x_1^2 + a_{15}x_1x_2 + a_{16}x_2^2 \\ a_{21}x_2 + a_{22}x_1x_2 + a_{23}x_2^2 + f_w(x_1) \end{bmatrix}, \quad (3.6)$$

where A is a matrix of unknown coefficients corresponding to a dictionary $\varphi(x)$ containing polynomial interactions up to second order, and $f_w(x_1)$ is a fully connected deep neural network with two hidden layers of dimension 20, a hyperbolic tangent activation and a set of unknown weight and bias parameters denoted by w . Note that the neural network admits only x_1 as an input to make sure that the resulting parametrized dynamical system has a unique solution. Under this set-up, the proposed Bayesian framework can be employed to jointly infer the dictionary and the neural network parameters that define the model representation, namely $\theta := \{a_{11}, a_{12}, a_{13}, a_{14}, a_{15}, a_{16}, a_{21}, a_{22}, a_{23}, w\}$. This defines a 490-dimensional inference problem.

Figure 10 shows the scattered training data, the exact simulated trajectory, the predicted trajectory corresponding to the MAP estimate of the inferred model parameters as well as a two-standard-deviation band around a set of plausible forecasted states. It is evident that the revised hybrid model formulation can now correctly identify the true underlying dynamics, leading to an accurate predicted MAP trajectory, while the predicted uncertainty effectively diminishes and concentrates around the ground truth. Moreover, the inferred coefficients of the irrelevant terms are very close to zero, while the active terms are all properly identified. This is also true for the neural network approximation of the missing closure term $-\beta \sin(x_1)$, as depicted in figure 11, which also includes uncertainty estimates over the neural network outputs. This simple example illustrates the great flexibility of the proposed Bayesian framework in seamlessly distilling parsimonious and interpretable models from imperfect data via physics-informed dictionary learning, as well harnessing the power of black-box representations for approximating missing closure terms.

(d) Bayesian calibration of a yeast glycolysis model

In this final example, the performance of the proposed algorithms applied to a realistic problem in systems biology is investigated. To this end, a yeast glycolysis process is considered and described

Figure 12. Yeast glycolysis dynamics: uncertainty estimation of the inferred model parameters obtained using the proposed BDP method. Estimates for the minimum, maximum, median, first quartile and third quartile are provided, while the true parameter values are highlighted in blue. (Online version in colour.)

by a seven-dimensional dynamical system [1,52] as

$$\left. \begin{aligned}
 \frac{dS_1}{dt} &= J_0 - k_1 S_1 A_3 \left[1 + \left(\frac{A_3}{K_I} \right)^q \right]^{-1}, \\
 \frac{dS_2}{dt} &= 2k_1 S_1 A_3 \left[1 + \left(\frac{A_3}{K_I} \right)^q \right]^{-1} - k_2 S_2 N_1 - k_6 S_2 N_2, \\
 \frac{dS_3}{dt} &= k_2 S_2 N_1 - k_3 S_3 A_2, \\
 \frac{dS_4}{dt} &= k_3 S_3 A_2 - k_4 S_4 N_2 - \kappa (S_4 - S_4^{\text{ex}}), \\
 \frac{dN_2}{dt} &= k_2 S_2 N_1 - k_4 S_4 N_2 - k_6 S_2 N_2, \\
 \frac{dA_3}{dt} &= -2k_1 S_1 A_3 \left[1 + \left(\frac{A_3}{K_I} \right)^q \right]^{-1} + 2k_3 S_3 A_2 - k_5 A_3 \\
 \text{and} \quad \frac{dS_4^{\text{ex}}}{dt} &= \phi \kappa (S_4 - S_4^{\text{ex}}) - k S_4^{\text{ex}},
 \end{aligned} \right\} \quad (3.7)$$

where $N_1 + N_2 = N$ and $A_2 + A_3 = A$. Here, it is clear that the system dynamics cannot be easily parametrized by a dictionary owing to the q exponent appearing as the order of the polynomial in the denominator. In this case, a straightforward application of dictionary learning approaches such as SINDy [17] would face fundamental difficulties. To this end, here we will assume that the model form is known, as prescribed by domain knowledge from systems biology, and our goal is to: (i) compute the posterior distribution over the unknown model parameters $\theta := \{J_0, k_1, k_2, k_3, k_4, k_5, k_6, k, \kappa, q, K_I, \phi, N, A\}$ from a small set of noisy and irregularly sampled observations and (ii) test the ability of the inferred model to accurately generalize from different initial conditions that were not observed during model training.

Figure 13. Learning yeast glycolysis dynamics from noisy data: future forecasts with quantified uncertainty from a previously unseen initial condition (i.e. an initial condition that was not used during model training). (Online version in colour.)

Table 6. Yeast glycolysis dynamics: MAP estimation of the inferred model parameters using noisy training data.

J_0	k_1	k_2	k_3	k_4	k_5	k_6	k	κ	q	K_I	ϕ	N	A
2.52	98.92	6.43	16.12	101.24	1.28	12.13	1.82	13.12	4.01	0.52	0.01	0.95	4.01

A training dataset is generated by randomly sub-sampling $n = 1000$ irregularly distributed observations from a single simulated trajectory of the system from an initial condition: $(S_1, S_2, S_3, S_4, N_2, A_3, S_4^{\text{ex}}) = (0.5, 1.9, 0.18, 0.15, 0.16, 0.1, 0.064)$ in the time interval $t \in [0, 5]$, assuming a ground truth set of parameters obtained from the experimental data provided in [1]: $J_0 = 2.5 \text{ mM min}^{-1}$, $k_1 = 100.0 \text{ mM min}^{-1}$, $k_2 = 6.0 \text{ mM min}^{-1}$, $k_3 = 16.0 \text{ mM min}^{-1}$, $k_4 = 100.0 \text{ mM min}^{-1}$, $k_5 = 1.28 \text{ min}^{-1}$, $k_6 = 12.0 \text{ mM min}^{-1}$, $k = 1.8 \text{ min}^{-1}$, $\kappa = 13.0 \text{ min}^{-1}$, $q = 4.0$, $K_I = 0.52 \text{ mM}$, $N = 1.0 \text{ mM}$, $A = 4.0 \text{ mM}$ and $\phi = 0.1$. Moreover, the training data are perturbed by a 2% white noise proportional to their standard deviations.

Table 6 summarizes the inferred MAP estimators for the unknown model parameters. Based on these results, all inferred parameters closely agree with the ground truth values used to generate the training data as reported in [1]. Uncertainty estimates for the inferred parameters can also be deduced from the computed posterior distribution $p(\theta|\mathcal{D})$, as presented in the box plots of figure 12, where the minimum, maximum, median, first quantile and third quantile obtained from the HMC simulations for each parameter are presented. Finally, note that all true values fall between the predicted quantiles, while the MAP estimators of all the parameters have considerably small relative errors than the true parameter values, as shown in table 6.

Finally, to illustrate the generalization capability of the inferred model with respect to different initial conditions from those used during training, the quality of the predicted states is assessed considering a random initial condition of $[0.428, 1.42, 0.11, 0.296, 0.252, 0.830, 0.064]$, which has not been observed during model training. The close agreement with the reference solution (figure 13) indicates that the inferred model is very capable of generalizing in terms of both handling different initial conditions and extrapolating to reliable future forecasts with quantified uncertainty.

4. Conclusion

We have presented a novel machine learning framework for robust systems identification under uncertainty. The proposed framework leverages state-of-the-art differential programming techniques in combination with gradient-enhanced sampling schemes for scalable Bayesian inference of high-dimensional posterior distributions to infer interpretable and parsimonious representations of complex dynamical systems from *imperfect* (e.g. sparse, noisy, irregularly sampled) data. The developed methods are general as they can seamlessly combine dictionary learning, domain knowledge and black-box approximations, all in a computationally efficient workflow with end-to-end uncertainty quantification. The effectiveness of the proposed techniques has been systematically investigated and compared with state-of-the-art approaches across a collection of prototype problems.

Although the proposed BDP framework provides great flexibility to infer a distribution over plausible parsimonious representations of a dynamical system, a number of technical issues need to be further investigated. The first relates to devising more effective initialization procedures for Markov chain Monte Carlo sampling. Here, we have partially addressed this via the MAP preconditioning algorithm discussed in §2e; however, during the preconditioning process, since the form of the dynamical system is unknown, the intermediate estimations of the parameters may cause the system to become stiff. Moreover, for cases where only very sparse observations are available, the model needs to be integrated with a large time step dt and stiffness of the system can lead to numerical instabilities during model training. A possible enhancement in this direction is to use more general stiffly stable ODE solvers as discussed in [14,15] or more sophisticated time-step annealing strategies [27]. Another potential future work could be identifying the uncertainty in the model's parameters with partial observations, such that some variables of the system are not accessible. Such a task would involve physics-informed regularization on the unknown latent dynamics of the system. Approaches used in [52] could be helpful for solving this problem. A third open question is how to adapt the proposed method to stochastic dynamical systems where the dynamics itself may be driven by a stochastic process. Approaches mentioned in [59] could be useful. Finally, the proposed BDP framework can be extended to parameter identification for partial differential equations (PDEs). The latter generally translate into high-dimensional dynamical systems after discretization. The learning task in this context could be carried out not only for the PDEs' parameters but also for the discretization scheme [14].

Although a series of promising empirical results were presented, we would like to also point out that a rigorous theoretical justification on the convergence of the proposed methods is still missing. To our understanding, there are two fundamental concepts that need to be investigated: (i) under which conditions a dynamical system is identifiable from raw trajectory data and (ii) the consistency of Bayesian inference (e.g. asymptotic behaviour as more data are collected,

posterior contraction rates, etc.). For the first point, and under the deterministic setting, theoretical results in compressed sensing [60] and sparse regression can provide a path to obtain theoretical guarantees. For example, the work of Brunton *et al.* [17] discusses the effect of observation noise on the identifiability of a dynamical system [61,62]. For the latter point, one could gain theoretical insight from analysing the properties of HMC samplers as discussed in [35,63]; in practice, we may only rely on appropriate tests of statistical convergence such as the Geweke analysis and the Gelman–Rubin tests [45,64], similarly to the results presented in §3a(iii).

Data accessibility. All code and data accompanying this manuscript are available at <https://github.com/PredictiveIntelligenceLab/BayesianDifferentiableProgramming>.

Authors' contributions. P.P. and Y.Y. conceived the methods, Y.Y. implemented the methods, Y.Y. and M.A.B. performed the simulations. Y.Y., M.A.B. and P.P. wrote the manuscript.

Competing interests. We declare we have no competing interests.

Funding. This work received support from the US Department of Energy under the Advanced Scientific Computing Research program (grant no. DE-SC0019116) and the Defense Advanced Research Projects Agency under the Physics of Artificial Intelligence program (grant no. HR00111890034) and the Air Force Office of Scientific Research (grant no. FA9550-20-1-0060).

References

1. Ruoff P, Christensen MK, Wolf J, Heinrich R. 2003 Temperature dependency and temperature compensation in a model of yeast glycolytic oscillations. *Biophys. Chem.* **106**, 179–192. (doi:10.1016/S0301-4622(03)00191-1)
2. Kak AC, Slaney M, Wang G. 2002 Principles of computerized tomographic imaging. *Med. Phys.* **29**, 107–107. (doi:10.1118/1.1455742)
3. Raissi M, Yazdani A, Karniadakis GE. 2020 Hidden fluid mechanics: learning velocity and pressure fields from flow visualizations. *Science* **367**, 1026–1030. (doi:10.1126/science.aaw4741)
4. Palmer TN. 1999 A nonlinear dynamical perspective on climate prediction. *J. Clim.* **12**, 575–591. (doi:10.1175/1520-0442(1999)012<0575:ANDPOC>2.0.CO;2)
5. Feinberg M, Horn FJ. 1974 Dynamics of open chemical systems and the algebraic structure of the underlying reaction network. *Chem. Eng. Sci.* **29**, 775–787. (doi:10.1016/0009-2509(74)80195-8)
6. Haller G. 2002 Lagrangian coherent structures from approximate velocity data. *Phys. Fluids* **14**, 1851–1861. (doi:10.1063/1.1477449)
7. Tantet A, Lucarini V, Lunkeit F, Dijkstra HA. 2018 Crisis of the chaotic attractor of a climate model: a transfer operator approach. *Nonlinearity* **31**, 2221. (doi:10.1088/1361-6544/aaaf42)
8. Bemporad A, Morari M. 1999 Control of systems integrating logic, dynamics, and constraints. *Automatica* **35**, 407–427. (doi:10.1016/S0005-1098(98)00178-2)
9. Lu F, Zhong M, Tang S, Maggioni M. 2019 Nonparametric inference of interaction laws in systems of agents from trajectory data. *Proc. Natl Acad. Sci. USA* **116**, 14 424–14 433. (doi: 10.1073/pnas.1822012116)
10. Krizhevsky A, Sutskever I, Hinton GE. 2012 Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, vol. 25 (eds F Pereira, CJC Burges, L Bottou, KQ Weinberger), pp. 1097–1105. Curran Associates, Inc.
11. Bahdanau D, Cho K, Bengio Y. 2014 Neural machine translation by jointly learning to align and translate. (<http://arxiv.org/abs/1409.0473>)
12. Yang Y, Perdikaris P. 2019 Conditional deep surrogate models for stochastic, high-dimensional, and multi-fidelity systems. *Comput. Mech.* **64**, 417–434. (doi:10.1007/s00466-019-01718-y)
13. Han J, Jentzen A, Weinan E. 2018 Solving high-dimensional partial differential equations using deep learning. *Proc. Natl Acad. Sci. USA* **115**, 8505–8510. (doi:10.1073/pnas.1718942115)
14. Rackauckas C, Ma Y, Martensen J, Warner C, Zubov K, Supekar R, Skinner D, Ramadhan A. 2020 Universal differential equations for scientific machine learning. (<http://arxiv.org/abs/2001.04385>)
15. Gholami A, Keutzer K, Biros G. 2019 Anode: unconditionally accurate memory-efficient gradients for neural odes. (<http://arxiv.org/abs/1902.10298>)
16. Chen RTQ, Rubanova Y, Bettencourt J, Duvenaud DK. 2018 Neural ordinary differential equations. In *Advances in neural information processing systems*, vol. 31 (eds S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, R Garnett), pp. 6571–6583. Curran Associates, Inc.

17. Brunton SL, Proctor JL, Kutz JN. 2016 Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. Natl Acad. Sci. USA* **113**, 3932–3937. (doi:10.1073/pnas.1517384113)
18. Rudy SH, Brunton SL, Proctor JL, Kutz JN. 2017 Data-driven discovery of partial differential equations. *Sci. Adv.* **3**, e1602614. (doi:10.1126/sciadv.1602614)
19. Brennan C, Venturi D. 2018 Data-driven closures for stochastic dynamical systems. *J. Comput. Phys.* **372**, 281–298. (doi:10.1016/j.jcp.2018.06.038)
20. Raissi M, Perdikaris P, Karniadakis GE. 2018 Multistep neural networks for data-driven discovery of nonlinear dynamical systems. (<http://arxiv.org/abs/1801.01236>)
21. Qin T, Wu K, Xiu D. 2019 Data driven governing equations approximation using deep neural networks. *J. Comput. Phys.* **395**, 620–635. (doi:10.1016/j.jcp.2019.06.042)
22. Raissi M, Perdikaris P, Karniadakis GE. 2017 Inferring solutions of differential equations using noisy multi-fidelity data. *J. Comput. Phys.* **335**, 736–746. (doi:10.1016/j.jcp.2017.01.060)
23. Raissi M, Perdikaris P, Karniadakis GE. 2019 Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* **378**, 686–707. (doi:10.1016/j.jcp.2018.10.045)
24. Zhu Y, Zabaras N, Koutsourelakis PS, Perdikaris P. 2019 Physics-constrained deep learning for high-dimensional surrogate modeling and uncertainty quantification without labeled data. *J. Comput. Phys.* **394**, 56–81. (doi:10.1016/j.jcp.2019.05.024)
25. Yang Y, Perdikaris P. 2019 Adversarial uncertainty quantification in physics-informed neural networks. *J. Comput. Phys.* **394**, 136–152. (doi:10.1016/j.jcp.2019.05.027)
26. Yang L, Zhang D, Karniadakis GE. 2020 Physics-informed generative adversarial networks for stochastic differential equations. *SIAM J. Sci. Comput.* **42**, A292–A317. (doi:10.1137/18M1225409)
27. Wang S, Teng Y, Perdikaris P. 2020 Understanding and mitigating gradient pathologies in physics-informed neural networks. (<http://arxiv.org/abs/2001.04536>)
28. Wang Q, Hesthaven JS, Ray D. 2019 Non-intrusive reduced order modeling of unsteady flows using artificial neural networks with application to a combustion problem. *J. Comput. Phys.* **384**, 289–307. (doi:10.1016/j.jcp.2019.01.031)
29. Deveney T, Mueller E, Shardlow T. 2019 A deep surrogate approach to efficient Bayesian inversion in PDE and integral equation models. (<http://arxiv.org/abs/1910.01547>)
30. Raissi M, Karniadakis GE. 2018 Hidden physics models: machine learning of nonlinear partial differential equations. *J. Comput. Phys.* **357**, 125–141. (doi:10.1016/j.jcp.2017.11.039)
31. Tartakovsky AM, Marrero CO, Perdikaris P, Tartakovsky GD, Barajas-Solano D. 2018 Learning parameters and constitutive relationships with physics informed deep neural networks. (<http://arxiv.org/abs/1808.03398>)
32. Kissas G, Yang Y, Hwuang E, Witschey WR, Detre JA, Perdikaris P. 2020 Machine learning in cardiovascular flows modeling: predicting arterial blood pressure from non-invasive 4D flow MRI data using physics-informed neural networks. *Comput. Methods Appl. Mech. Eng.* **358**, 112623. (doi:10.1016/j.cma.2019.112623)
33. Chen Y, Lu L, Karniadakis GE, Negro LD. 2019 Physics-informed neural networks for inverse problems in nano-optics and metamaterials. (<http://arxiv.org/abs/1912.01085>)
34. Sahli Costabal F, Yang Y, Perdikaris P, Hurtado DE, Kuhl E. 2020 Physics-informed neural networks for cardiac activation mapping. *Front. Phys.* **8**, 42. (doi:10.3389/fphy.2020.00042)
35. Neal RM *et al.* 2011 MCMC using Hamiltonian dynamics. In *Handbook of Markov chain Monte Carlo* (eds S Brooks, A Gelman, GL Jones, X-L Meng). Boca Raton, FL: CRC Press.
36. Pontryagin LS, Boltyanskii VG, Gamkrelidze RV, Mishchenko EF. 1962 *The mathematical theory of optimal processes*. New York, NY: Wiley.
37. van Merriënboer B, Breuleux O, Bergeron A, Lamblin P. 2018 Automatic differentiation in ML: where we are and where we should be going. In *Advances in neural information processing systems*, vol. 31 (eds S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, R Garnett), pp. 8757–8767. Curran Associates, Inc.
38. Hastie T, Tibshirani R, Friedman J. 2009 *The elements of statistical learning: data mining, inference, and prediction*. New York, NY: Springer Science & Business Media.
39. Ghahramani Z. 2015 Probabilistic machine learning and artificial intelligence. *Nature* **521**, 452–459. (doi:10.1038/nature14541)
40. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC. 2017 Improved training of Wasserstein GANs. In *Advances in neural information processing systems*, vol. 30 (eds I Guyon,

- UV Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, R Garnett, pp. 5767–5777. Curran Associates, Inc.
41. Su W, Boyd S, Candes E. 2014 A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. In *Advances in neural information processing systems*, vol. 27 (eds Z Ghahramani, M Welling, C Cortes, N Lawrence, KQ Weinberger), pp. 2510–2518. Curran Associates, Inc.
 42. Bottou L. 2010 Large-scale machine learning with stochastic gradient descent. In *Proc. of COMPSTAT'2010* (eds Y Lechevallier, G Saporta), pp. 177–186. Heidelberg, Germany: Physica-Verlag HD.
 43. Iserles A. 2008 *A first course in the numerical analysis of differential equations*. Cambridge Texts in Applied Mathematics, 2nd edn. Cambridge, UK: Cambridge University Press.
 44. Geweke J. 1993 Bayesian treatment of the independent Student-t linear model. *J. Appl. Econ.* **8**, S19–S40. (doi:10.1002/jae.3950080504)
 45. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. 2013 *Bayesian data analysis*. Boca Raton, FL: CRC Press.
 46. Winkler RL. 1967 The assessment of prior distributions in Bayesian analysis. *J. Am. Stat. Assoc.* **62**, 776–800. (doi:10.1080/01621459.1967.10500894)
 47. Bernardo JM. 1979 Reference posterior distributions for Bayesian inference. *J. R. Stat. Soc.: Ser. B (Methodological)* **41**, 113–128. (doi:10.1111/j.2517-6161.1979.tb01066.x)
 48. Berger JO. 1990 Robust Bayesian analysis: sensitivity to the prior. *J. Stat. Plann. Inference* **25**, 303–328. (doi:10.1016/0378-3758(90)90079-A)
 49. Williams PM. 1995 Bayesian regularization and pruning using a Laplace prior. *Neural Comput.* **7**, 117–143. (doi:10.1162/neco.1995.7.1.117)
 50. Gilks WR, Best NG, Tan K. 1995 Adaptive rejection Metropolis sampling within Gibbs sampling. *J. R. Stat. Soc.: Ser. C (Applied Statistics)* **44**, 455–472. (doi:10.2307/2986138)
 51. Millar RB, Meyer R. 2000 Non-linear state space modelling of fisheries biomass dynamics by using Metropolis–Hastings within-Gibbs sampling. *J. R. Stat. Soc.: Ser. C (Applied Statistics)* **49**, 327–342. (doi:10.1111/1467-9876.00195)
 52. Yazdani A, Lu L, Raissi M, Karniadakis GE. 2020 Systems biology informed deep learning for inferring parameters and hidden dynamics. *PLOS Comput. Biol.* **16**. (doi:10.1371/journal.pcbi.1007575)
 53. Champion K, Lusch B, Kutz JN, Brunton SL. 2019 Data-driven discovery of coordinates and governing equations. *Proc. Natl Acad. Sci. USA* **116**, 22 445–22 451. (doi:10.1073/pnas.1906995116)
 54. Schnakenberg J. 1979 Simple chemical reaction systems with limit cycle behaviour. *J. Theor. Biol.* **81**, 389–400. (doi:10.1016/0022-5193(79)90042-0)
 55. Glorot X, Bengio Y. 2010 Understanding the difficulty of training deep feedforward neural networks. In *Proc. of the 13th Int. Conf. on Artificial Intelligence and Statistics, Sardinia, Italy, 13–15 May 2010* (eds YW Teh, M Titterton), pp. 249–256. JMLR Workshop and Conference Proceedings.
 56. Kingma DP, Ba J. 2014 Adam: a method for stochastic optimization. (<http://arxiv.org/abs/1412.6980>)
 57. Hoffman MD, Gelman A. 2014 The no-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* **15**, 1593–1623.
 58. Lusch B, Kutz JN, Brunton SL. 2018 Deep learning for universal linear embeddings of nonlinear dynamics. *Nat. Commun.* **9**, 4950. (doi:10.1038/s41467-018-07210-0)
 59. Li X, Wong TKL, Chen RT, Duvenaud D. 2020 Scalable gradients for stochastic differential equations. (<http://arxiv.org/abs/2001.01328>)
 60. Donoho DL. 2006 Compressed sensing. *IEEE Trans. Inf. Theory* **52**, 1289–1306. (doi:10.1109/TIT.2006.871582)
 61. Schmidt MD, Vallabhajosyula RR, Jenkins JW, Hood JE, Soni AS, Wikswo JP, Lipson H. 2011 Automated refinement and inference of analytical models for metabolic networks. *Phys. Biol.* **8**, 055011. (doi:10.1088/1478-3975/8/5/055011)
 62. Daniels BC, Nemenman I. 2015 Automated adaptive inference of phenomenological dynamical models. *Nat. Commun.* **6**, 1–8. (doi:10.1038/ncomms9133)
 63. Betancourt M, Byrne S, Livingstone S, Girolami M *et al.* 2017 The geometric foundations of Hamiltonian Monte Carlo. *Bernoulli* **23**, 2257–2298. (doi:10.3150/16-BEJ810)
 64. Cowles MK, Carlin BP. 1996 Markov chain Monte Carlo convergence diagnostics: a comparative review. *J. Am. Stat. Assoc.* **91**, 883–904. (doi:10.1080/01621459.1996.10476956)