

Neural ODEs with Irregular and Noisy Data

Pawan Goyal* Peter Benner[†]

*Max Planck Institute for Dynamics of Complex Technical Systems, 39106 Magdeburg, Germany.
Email: goyalp@mpi-magdeburg.mpg.de, ORCID: [0000-0003-3072-7780](https://orcid.org/0000-0003-3072-7780)

[†]Max Planck Institute for Dynamics of Complex Technical Systems, 39106 Magdeburg, Germany.
Email: benner@mpi-magdeburg.mpg.de, ORCID: [0000-0003-3362-4103](https://orcid.org/0000-0003-3362-4103)

Abstract: Measurement noise is an integral part while collecting data of a physical process. Thus, noise removal is necessary to draw conclusions from these data, and it often becomes essential to construct dynamical models using these data. We discuss a methodology to learn differential equation(s) using noisy and irregular sampled measurements. In our methodology, the main innovation can be seen in the integration of deep neural networks with the neural ordinary differential equations (ODEs) approach. Precisely, we aim at learning a neural network that provides (approximately) an implicit representation of the data and an additional neural network that models the vector fields of the dependent variables. We combine these two networks by constraining using neural ODEs. The proposed framework to learn a model describing the vector field is highly effective under noisy measurements. The approach can handle scenarios where dependent variables are not available at the same temporal grid. Moreover, a particular structure, e.g., second-order with respect to time, can easily be incorporated. We demonstrate the effectiveness of the proposed method for learning models using data obtained from various differential equations and present a comparison with the neural ODE method that does not make any special treatment to noise.

Keywords: Machine learning, deep neural networks, nonlinear differential equations, neural ODEs, second-order systems, irregular data

Novelty statement:

- This work blends neural networks with the neural ODE methods to learn dynamical models from noisy measurements, as well as to obtain denoised data.
- More precisely, two networks, one for an implicit representation of data and the second one for describing dynamics, are combined by an integral form of ODEs.
- An extension to second-order ODEs is discussed.
- The proposed methodology is capable of a situation when the dependent variables are sampled irregularly. Moreover, they need not be measured on the same irregular time grid.

Code link: https://gitlab.mpi-magdeburg.mpg.de/goyalp/implicit_neuralodes

1. Introduction

Uncovering dynamical models explaining physical phenomena and dynamic behaviors has been active research for centuries ¹. When a model describing the underlying dynamics is available, it can be used for several engineering studies such as process design, optimization, predictions, and control. Conventional approaches

¹For example, Isaac Newton developed his fundamental laws based on measured data.

based on physical laws and empirical knowledge are often used to derive dynamical models. However, this is impenetrable for many complex systems, e.g., understanding the Arctic ice pack dynamics, sea ice, power grids, neuroscience, or finance, to only name a few applications. Data-driven methods to discover models have enormous potential to understand transient behaviors in the latter cases better. Furthermore, data acquired using imaging devices or sensors are contaminated with measurement noise. Therefore, systematic approaches that learn a dynamic model with proper treatment of noise are required.

Towards this aim, the initial work [1] proposes a framework that explicitly incorporates the noise into a numerical time-stepping method, namely a *Runge-Kutta* method. Though the approach has shown promising directions, its scalability remains ambiguous as the approach explicitly needs noise estimates and aims to decompose the signal explicitly into noise and ground truth. Moreover, it requires that the Runge-Kutta method can give a reasonable estimate at the next step. Additionally, irregular sampling (e.g., when dependent variables are not even collected at the same time-grid) cannot be applied, which can be highly relevant when information is gathered from various sources, e.g., medical applications. This work discusses a deep learning-based approach to learning a dynamic model by attenuating neural networks with adaptive numerical integrations. It allows learning models to represent the vector field accurately without estimating noise explicitly and when dependent variables are arbitrarily irregularly sampled.

Our contributions: Our work introduces a framework to learn dynamics models by innovatively blending neural networks and numerical integration methods from noisy and irregular measurements. Precisely, we aim at learning two networks; one that approximately represents given measurement data implicitly, and the second one approximates the vector field. We connect these two networks by enforcing an integral form of ODEs as depicted in Figure 1. The appeal of the approach is that we do not require an explicit estimate of noise to learn a model. Furthermore, the proposed approach is applicable even if each dependent variable is collected on a different time grid, which can be irregular.

The remaining structure of the paper is as follows. In the next section, we present a summary of relevant work. In Section 3, we present our deep learning-based framework for learning dynamics from noisy measurements by combining two networks. There, we also demonstrate the effectiveness of the proposed methodology using various synthetic data with increasing noise levels. Section 5 discusses the application of learning second-order dynamical models. Moreover, in Section 6, we discuss how to handle irregular sampling of measurements. We conclude the paper with a summary and future research directions.

2. Relevant work

Data-driven methods to learn dynamic models have been studied for several decades, see, e.g., [2–4]. Learning linear models from input-output data goes back to Ho and Kalman [5]. There have been several algorithmic developments for linear systems, for example, the eigensystem realization algorithm [6, 7], and Kalman filter-based approaches [8–10]. Dynamic mode decomposition has also emerged as a promising approach to construct models from input-output data and has been widely applied in fluid dynamics applications, see, e.g., [11–13]. Furthermore, there has been a series of developments to learn nonlinear dynamic models. This includes, for example, equations free modeling [14], nonlinear regression [15], dynamic modeling [16], and automated inference of dynamics [17–19]. Utilizing symbolic regression and an evolutionary algorithm [20, 21], learning compact nonlinear models becomes possible. Moreover, leveraging sparsity (also known as sparse regression), several approaches have been proposed [22–27]. We also mention the work [28] that learns models using Gaussian process regression. All these methods have particular approaches to handle noise in the data. For example, sparse regression methods, e.g., [22, 23, 27] often utilize smoothing methods before identifying models, and the work [28] handles measurement noise as data represented like a Gaussian process.

Even though the aforementioned nonlinear modeling methods are appealing and powerful in providing analytic expressions for models, they are often built upon model hypotheses. For example, the success of sparse regression techniques relies on the fact that the nonlinear basis functions, describing the dynamics, lie in a candidate features library. For many complex dynamics, such as the melting Arctic ice, the utilization of these methods is not trivial. Thus, machine learning techniques, particularly deep learning-based ones, have emerged as powerful methods capable of expressing any complex function in a black-box manner given enough training data. Neural network-based approaches in the context of dynamical systems have been discussed in

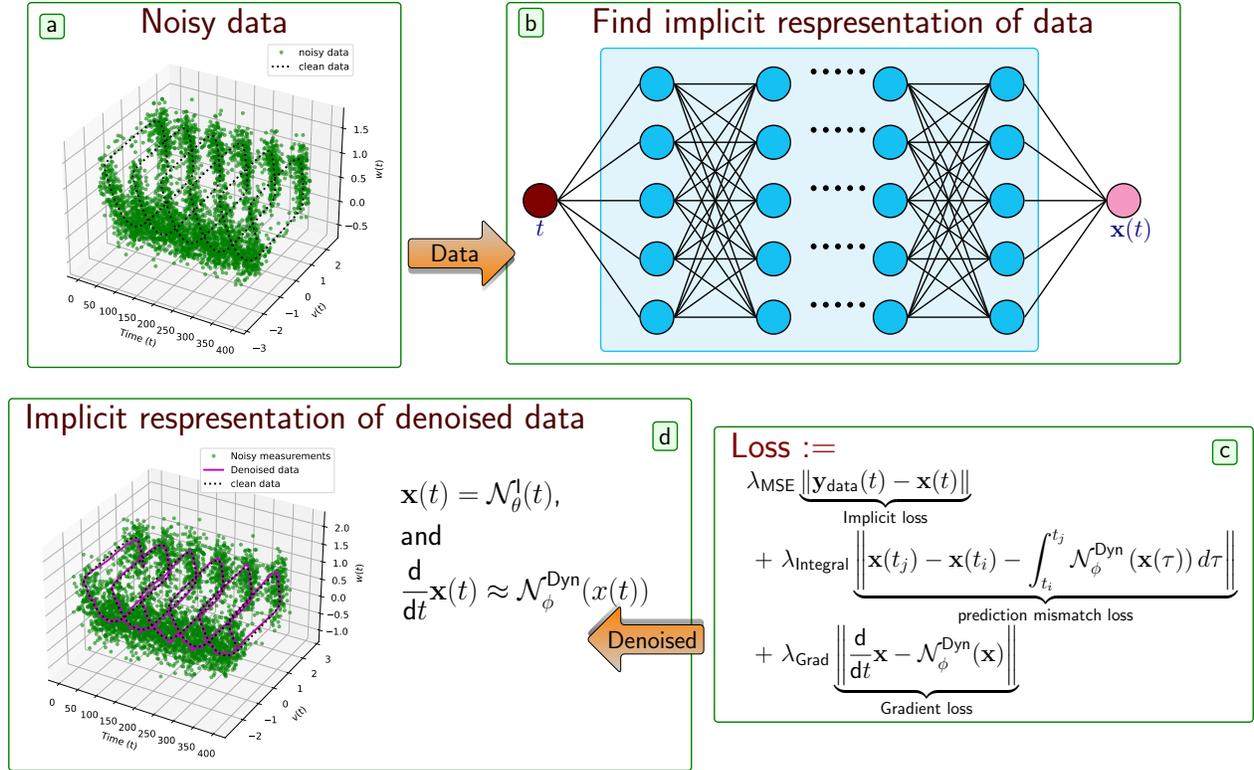


Figure 1.: The figure illustrates the framework for de-noise data and learning a model describing underlying dynamics. For this, we determine an implicit representation of data (approximately) by a network \mathcal{N}_{θ}^I and another network for the vector field $\mathcal{N}_{\phi}^{\text{Dyn}}$. These two networks are connected by enforcing that the dynamics of the output of the implicit representation can be given by $\mathcal{N}_{\phi}^{\text{Dyn}}$. Once the loss is minimized (shown in (c)), we obtain an implicit network for de-noised data and a model for the vector field $\mathcal{N}_{\phi}^{\text{Dyn}}$.

[29–32] decades ago. A particular type of neural network, namely recurrent neural networks, intrinsically models sequences and is often used for forecasting [33–37] but does not explicitly learn the corresponding vector field. Deep learning is also utilized to identify a coordinate transformation so that the dynamics in the transformed coordinates are almost linear or sparse in a high-dimensional feature basis, see, e.g., [38–41]. Furthermore, we mention that classical numerical schemes are incorporated with feed-forward neural networks to have discrete-time steppers for predictions, see [31, 42–44]. The approaches in [31, 42] can be interpreted as nonlinear autoregressive models [4]. A crucial feature of deep learning-based approaches that integrates numerical integration schemes is that vector fields are estimated using neural networks. Also, time-stepping is done using a numerical integration scheme. Furthermore, in recent times, *neural ordinary differential equations* (Neural ODEs) in which neural networks define the vector fields, have been proposed in [45], where it is shown how to compute gradient with respect to network parameters efficiently using adjoint sensitivities. As a result, one can utilize efficient black-box numerical solvers to solve ODEs in a given time span using adaptive time-stepping.

However, measurement data are often corrupted with noise, and these mentioned approaches do not perform any specific noise treatment. The work in [1] proposes a framework that explicitly incorporates the noise into a numerical time-stepping method. Though the approach has shown a promising direction, its scalability remains ambiguous. The approach explicitly needs noise estimates by learning the decomposition of the signal into noise and ground truth. Also, it relies on a Runge-Kutta scheme that can accurately estimate the variable at the next step.

Furthermore, in scenarios where the data are collected on an irregular time grid, the work [46] discussed a methodology by combining gated recurrent unit (GRU) and neural ODEs. In the approach, an estimate for

the initial condition of (latent) ODEs is learned, and an ODE for the vector field is then integrated using the estimated initial condition. However, long sequences are quite challenging to estimate the initial condition given measurements future in time. Although in [46], the measurements can be collected at an irregular time grid, it still requires that all dependent variables are measured at the same time grid. In the cases when each dependent variable is measured at a different time-grid, then the approach [46] is not applicable.

3. Proposed Methodology—Implicit Networks Combined Neural ODEs

In this section, we discuss our framework to learn dynamical models using noisy measurements without explicitly estimating noise. To achieve the goal, we utilize the powerful approximation capabilities of deep neural networks and their automatic differentiation feature with the neural ODEs approach [45], which allows integrating a function, defining the vector field, with any desired method and accuracy. For this, let us consider an autonomous nonlinear differential equation:

$$\frac{d}{dt}\mathbf{x}(t) = \mathbf{g}(\mathbf{x}(t)), \quad \mathbf{x}(0) = \mathbf{x}_0, \quad (1)$$

where $\mathbf{x}(t) \in \mathbb{R}^n$ denotes the solution at time t , and the continuous function $\mathbf{g}(\cdot) : \mathbb{R}^n \rightarrow \mathbb{R}^n$ defines the vector field. Furthermore, the solution $\mathbf{x}(t_j)$ can be explicitly given as:

$$\mathbf{x}(t_j) = \mathbf{x}(t_i) + \int_{t_i}^{t_j} \mathbf{g}(\mathbf{x}(\tau))d\tau. \quad (2)$$

Next, we further proceed to discuss our framework to learn dynamical models from noisy measurements. The approach involves two networks. The first network implicitly represents the variable as shown in Figure 1(b), and the second network approximates the vector field, or the function $\mathbf{g}(\cdot)$. These two networks are related by connecting the dependent variables at time t_i and t_j as shown in (2). That is, the output of the implicit network is not only in the vicinity of the measurement data but also its time-evolution can be defined by $\mathbf{g}(\mathbf{x})$ or (2). To be mathematically precise, let us denote noisy measurement data at time t_i by $\mathbf{y}(t_i)$. Furthermore, we consider a feed-forward neural network, denoted by \mathcal{N}_θ^I parameterized by θ , that approximately yields an implicit representation of measurement data, i.e.,

$$\mathcal{N}_\theta^I(t_i) := \mathbf{x}(t_i) \approx \mathbf{y}(t_i), \quad (3)$$

where $i \in \{1, \dots, m\}$ with m being the total number of measurements. Additionally, let us denote another neural network by $\mathcal{N}_\phi^{\text{Dyn}}$ parameterized by ϕ that approximates the vector field $\mathbf{g}(\cdot)$. We connect these two networks by enforcing that the time-evolution of the output of the network \mathcal{N}_θ^I can be described by $\mathcal{N}_\phi^{\text{Dyn}}$, i.e.,

$$\mathbf{x}(t_{i+1}) \approx \mathbf{x}(t_i) + \int_{t_i}^{t_{i+1}} \mathbf{g}(\mathbf{x}(\tau))d\tau, \quad \text{and} \quad \frac{d}{dt}\mathbf{x}(t_i) \approx \mathcal{N}_\phi^{\text{Dyn}}(\mathbf{x}(t_i)). \quad (4)$$

As a result, our goal becomes to determine the network parameters $\{\theta, \phi\}$ such that the following loss is minimized:

$$\mathcal{L} = \lambda_{\text{MSE}} \cdot \mathcal{L}_{\text{MSE}} + \lambda_{\text{Integral}} \cdot \mathcal{L}_{\text{Integral}} + \lambda_{\text{Grad}} \cdot \mathcal{L}_{\text{Grad}}, \quad (5)$$

where

- \mathcal{L}_{MSE} denotes the mean square error of the output of the network \mathcal{N}_θ^I and noisy measurements, i.e.,

$$\|\mathcal{N}_\theta^I(t_i) - \mathbf{y}(t_i)\|_F^2. \quad (6)$$

The loss enforces measurement data to be in the vicinity of the output of the implicit network, and λ_{MSE} is its weighting parameter.

- The term $\mathcal{L}_{\text{Integral}}$ links the two networks by comparing the prediction, i.e.,

$$\left\| \mathbf{x}(t_j) - \mathbf{x}(t_i) - \int_{t_i}^{t_j} \mathbf{g}(\mathbf{x}(\tau))d\tau \right\|_F^2, \quad (7)$$

and the parameter $\lambda_{\text{Integral}}$ defines its weight in the total loss.

- The vector field at the output of the implicit network can also be computed directly using automatic differentiation, but it also can be computed using the network $\mathcal{N}_\phi^{\text{Dyn}}$. The term $\mathcal{L}_{\text{Grad}}$ penalizes its mismatch as follows:

$$\left\| \mathcal{N}_\phi^{\text{Dyn}}(\mathbf{x}(t_i)) - \frac{d}{dt}\mathbf{x}(t_i) \right\|_F^2, \quad (8)$$

and λ_{Grad} is its corresponding regularization parameter.

The total loss \mathcal{L} can be minimized using a gradient-based optimizer such as Adam [47]. Once the networks are trained and have found their parameters that minimize the loss, we can generate the denoised variables using the implicit network \mathcal{N}_θ^I , and the vector field by the network $\mathcal{N}_\phi^{\text{Dyn}}$. In the rest of the paper, we denote the proposed methodology using Implicit-Neural ODEs (in sort **Imp-NODEs**).

4. Numerical Experiments

We now present the performance of the approach discussed in Section 3 to de-noise measurement data, as well as to learn a model for estimating the vector field by means of an example. To that aim, we consider data obtained by solving a differential equation that is then corrupted using white Gaussian noise by varying the noise level. For a given percentage, we determine the noise as follows:

$$\nu \sim \mathcal{N}(0, \mu), \quad \text{with } \mu = \text{Noise}\%.$$

We have implemented our framework using the deep learning library PyTorch [48] and have optimized both networks together using the Adam optimizer [47]. We have used `torchdiffeq` [45], a Python package to integrate an ODE and to do back-propagation to determine gradients, with the default settings.

Moreover, to obtain a good initial guess to employ the proposed **Imp-NODEs**, we replace the integration term by its approximation using 4th-order Runge-Kutta (RK4) method. It is computationally faster because the RK4 method takes the fixed four calls of the function, defining the vector field. We first train using this strategy for 5 000 epochs, followed by training using an adaptive ODE integration scheme within a given time span for 10 000 epochs for $\mu = \{1\%, 5\%\}$, and for 1000 epochs only for $\mu = \{10\%, \dots, 50\%\}$ to do early stopping to avoid over-fitting. We also make use of a learning scheduler, for which we reduced the learning rate by one-tenth after each 4 000 epoches. The neural network architecture design and hyperparameters are discussed in Appendix A, and we have run all our experiments on a NVIDIA[®] P100 GPU.

Cubic damped model: We consider a damped cubic system, which is described by

$$\begin{aligned} \dot{\mathbf{x}}(t) &= -0.1\mathbf{x}(t)^3 + 2.0\mathbf{y}(t)^3, \\ \dot{\mathbf{y}}(t) &= -2.0\mathbf{x}(t)^3 - 0.1\mathbf{y}(t)^3. \end{aligned} \quad (9)$$

It has been one of the benchmark examples in discovering models using data, see, e.g., [27, 49] but there, it is assumed that the dynamics can be given sparsely in a high-dimensional feature dictionary. Here, we do not make any such assumptions and instead learn the vector field using a neural network. For this example, we take 2 500 data points in the time interval $[0, 25]$ by simulating the model using the initial condition $[2, 0]$. We add various noise levels in the clean data to have noisy measurements synthetically. We construct neural networks for the implicit representation and the vector field with the parameters given in Table 1.

We corrupt the data by adding mean-zero Gaussian white noise of variances $\{1\%, \dots, 50\%\}$. We aim to obtain a denoised signal and a model, defining its vector field. Before employing the method, we perform a pre-processing step to noisy data using a low-pass filter to remove a large portion of the high-frequency noise. We compare our methodology with the neural ODE framework [45], which also focuses on learning a neural network, defining the underlying vector field.

To train the implicit network, and the neural network for ODEs, we set $\lambda_{\text{Integral}} = 1.0$ and $\lambda_{\text{Grad}} = 10^{-2}$ in the loss function (5), and choose $\lambda_{\text{MSE}} = 1.0$ for $\mu = \{1\%, 5\%\}$, and $\lambda_{\text{MSE}} = 0.5$ for $\mu = \{10\%, 20\%\}$, and $\lambda_{\text{MSE}} = 0.2$ for $\{30\%, 40\%, 50\%\}$ to avoid over-fitting of noisy data using the implicit network. Moreover, to integrate ODEs, we consider the time-span of $\text{bs} \cdot \text{dt}$, where **bs** is treated as a batch-size and is set to 4 and $\text{dt} = 10^{-2}$.

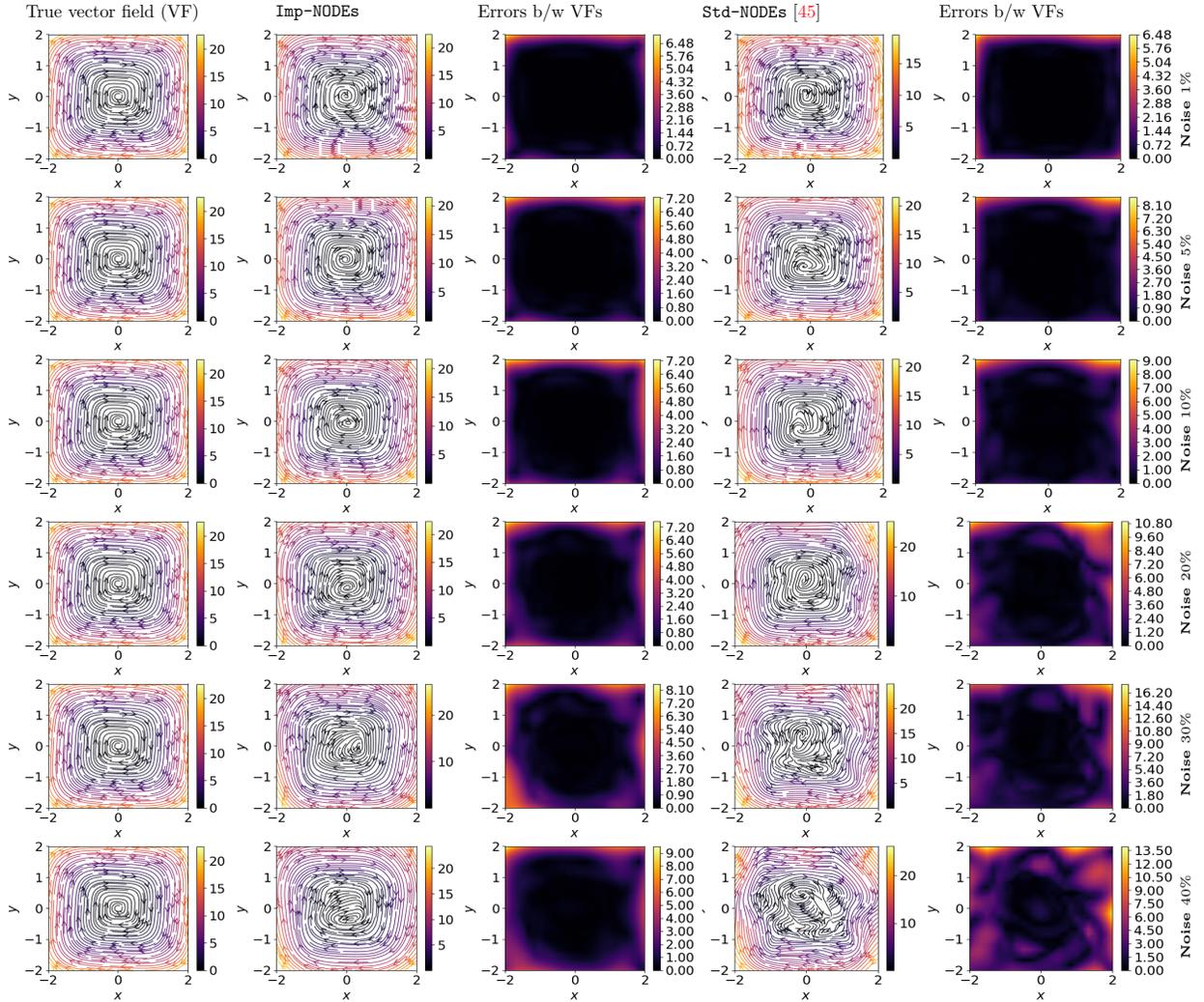


Figure 2.: Cubic2D example: A comparison of vector fields of the ground truth and learned models for various noise. The first column shows the ground truth vector field; the second and fourth columns show the learned vector fields from the Imp-NODEs and Std-NODEs, respectively; the third and fifth columns present the errors between the learned and the ground truth vector fields.

Having trained models, we plot the results in Figure 2, where the learned vector fields from the proposed method (Imp-NODEs), and it is compared with neural ODE [45] (Std-NODEs). Std-NODEs is also trained with the same configurations and the same number of epochs as Imp-NODEs. It is clear from the figures that Imp-NODEs is able to learn the underlying vector field faithfully, whereas the Std-NODEs fails to identify the vector field accurately. It is quite evident for higher levels of noise. Our approach consists of an implicit network, aiming to generate denoised data whose dynamics can be defined by a neural network. Thus, we plot the denoised data obtained from the implicit network in Figure 3. It shows that using the implicit network, we can obtain denoised data, close to the ground truth clean data even for a high noise level, which, otherwise, is not possible by employing solely Std-NODEs.

We further compute the mean and median errors between the learned vector fields and the ground truth for a quantitative comparison of these approaches to learning vector fields. For batch size $bs = 4$, we plot these errors in Figure 4a, showing the robustness of the proposed approach with respect to the noise levels. Furthermore, we report the performance of both approaches by changing the batch size, meaning by varying the time span for integration (for batch size bs , the time span is $bs \times dt$) for the corrupted data with 5% noise. For this study, we again plot the mean and medium errors between the learned and ground truth

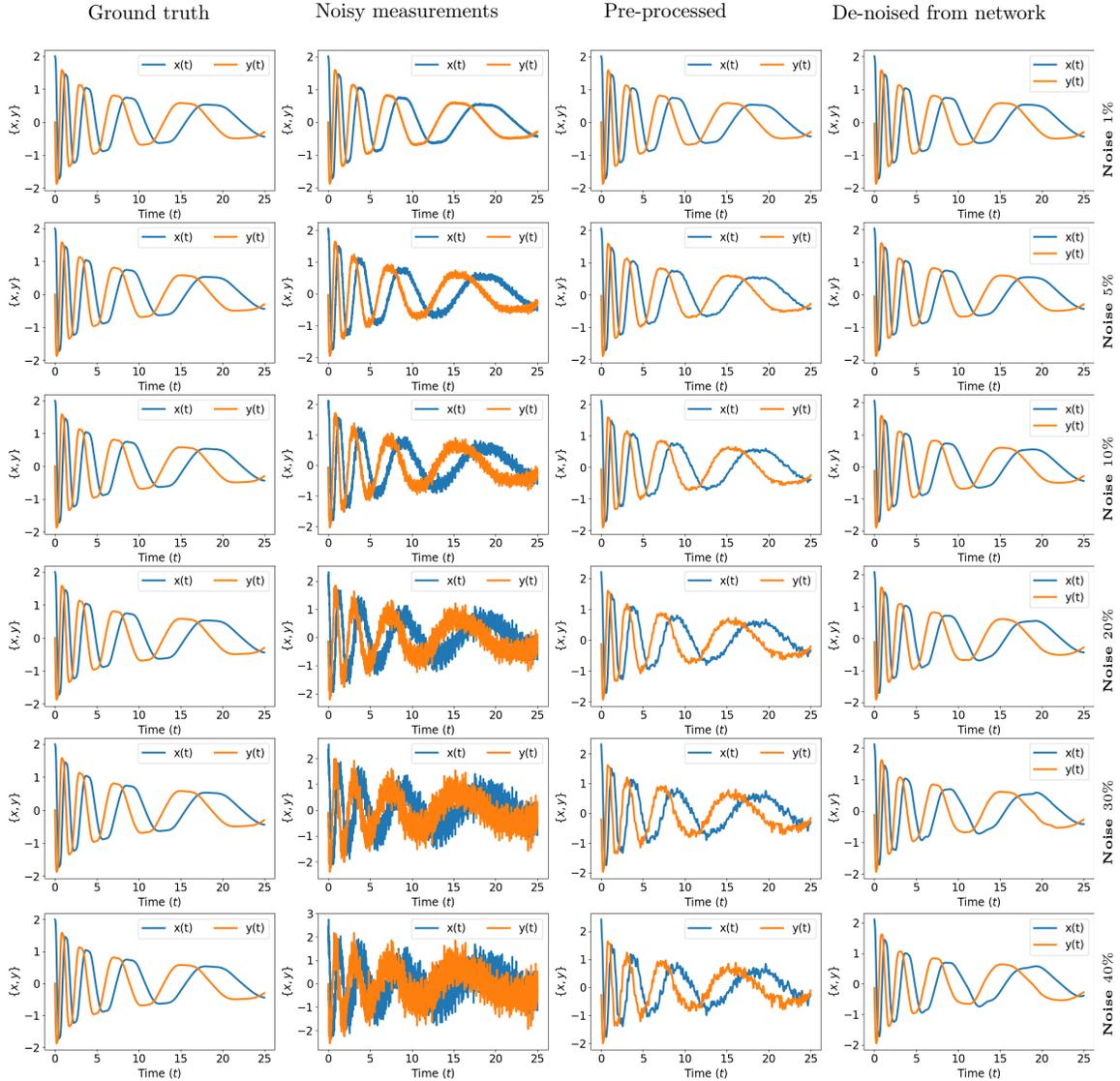


Figure 3.: Cubic2D example: Denoised data from the trained implicit network for various noise levels. The plots contain the ground truth, noisy measurements, pre-processed data using a low-pass filter, and denoised data from the implicit network.

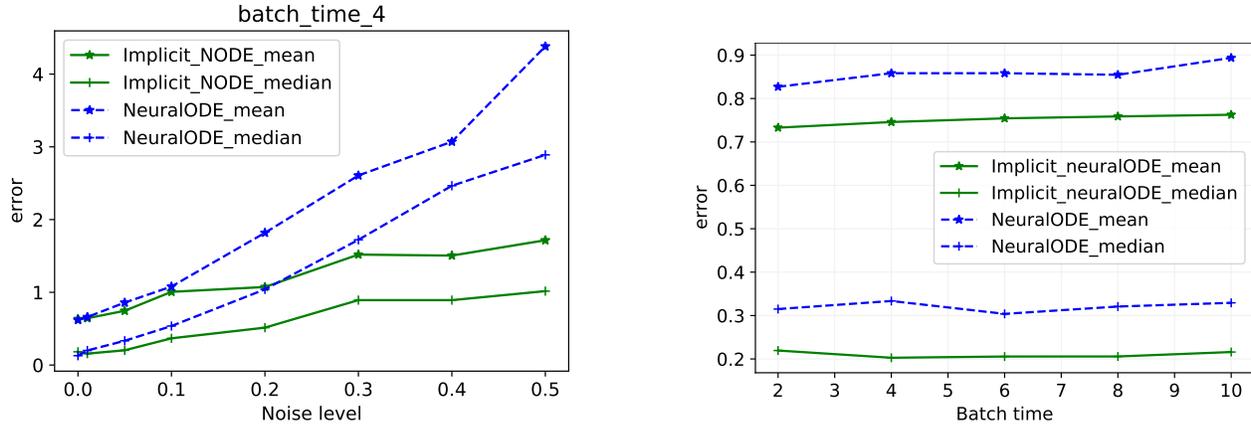
vector fields in Figure 4b. We notice that the errors are of the same order with respect to the batch size for **Imp-NODEs** and perform better than **Std-NODEs** in both mean and median errors for all batch sizes. We highlight that although the errors are of the same order with respect to **bs**, we notice a substantial increase in computational cost because of a large integration time span.

Based on these studies, we empirically conclude that an implicit network looks at generating data in the vicinity of the measurements, which can be integrated by an ODE defined by a neural network.

5. Second Order Neural ODEs for Noisy Measurements

Several dynamics of engineering process, particularly in electrical and mechanical ones are of second-order, which can be given as follows:

$$\ddot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), \dot{\mathbf{x}}(t)), \tag{10}$$



(a) Mean and median error of the vector fields for Imp-NODEs and Std-NODEs for various levels of noise.

(b) Mean and median error of the vector fields for Imp-NODEs and Std-NODEs for different batch size.

Figure 4.: Cubic2D example: Comparisons of the error between the learned and ground truth vector fields for different noise levels and for different batch sizes.

where $\dot{\mathbf{x}}(t)$ and $\ddot{\mathbf{x}}(t)$ denote the first and second derivatives of $\mathbf{x}(t)$ with respect to t , respectively. As discussed in [50], it is advantageous to consider the companion first-order system of (10) which is

$$\begin{bmatrix} \ddot{\mathbf{x}}(t) \\ \dot{\mathbf{x}}(t) \end{bmatrix} = \begin{bmatrix} \mathbf{f}(\mathbf{x}(t), \dot{\mathbf{x}}(t)) \\ \dot{\mathbf{x}}(t) \end{bmatrix}, \quad (11)$$

which inherently preserves the second-order behavior. The above system can be seen as a first-order system with a constraint. The method proposed in the previous section can be readily applied to learn second-order neural ODEs for noisy measurements by incorporating implicit networks.

Numerical example: Pendulum dynamics

To illustrate learning second-order dynamics, we consider a nonlinear pendulum model described as

$$\ddot{\mathbf{x}}(t) = -\sin(\mathbf{x}(t)) - 0.05 \cdot \dot{\mathbf{x}}(t). \quad (12)$$

We collect data using the initial condition $[\dot{\mathbf{x}}(t), \mathbf{x}(t)] = [-0.5, 2.0]$ at the time interval $dt = 0.25s$, which is then corrupted by adding Gaussian white noise of $\mu = \{5\%, 20\%\}$. We do not apply any pre-processing step for this example since filtering with a large time-step is not trivial using a simple filter such as a low-pass filter. Hence, we will also observe the filtering capability of the implicit network with the raw data to obtain denoised data. We employ the proposed scheme by combining an implicit network and neural ODEs by imposing the second-order structure. We train networks with parameters $\lambda_{\text{Integral}} = 1.0$, $\lambda_{\text{Grad}} = 10^{-2}$, and $\lambda_{\text{MSE}} = 1.0$ in (5) for $\mu = 5\%$, and $\lambda_{\text{MSE}} = 0.1$ for $\mu = 20\%$ to avoid over-fitting. We also use an early-stopping for over-fitting, as discussed in the previous example. For integration, we take the time-span of $2 \cdot dt$.

We compare our results with neural ODEs for second-order systems, the approach proposed in [50]. We plot the learned vector field from both methods in Figure 5, where we see better performance for the proposed method than the one proposed in [50]. Particularly, it is quite apparent for a larger noise, see Figure 5, where Std-NODEs for second-order systems fails to compare the vector field (second row). Moreover, we also plot the denoised data, which is the output of the trained implicit network in Figure 6, indicating recovery of the data faithfully without using any prior pre-processing step as such.

6. Measurements at Irregular Sampling

Lastly, we illustrate the ready applicability of the proposed method when the data are collected at an irregular time grid, especially when dependent variables are not even measured in the same time frame. It is of particular

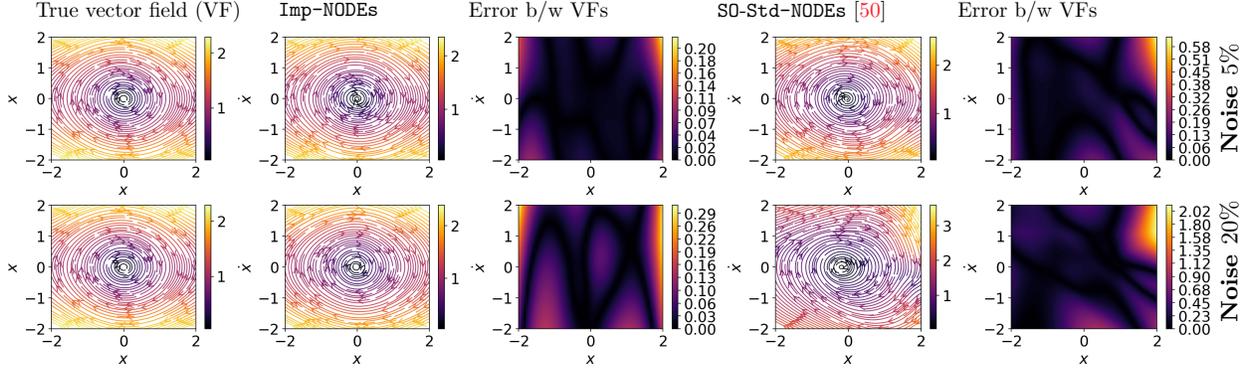


Figure 5.: Pendulum example: A comparison of the learned vector fields for second-order dynamical models using the proposed methodology and the one proposed in [50] for noise {5%, 20%}. The leftmost figure shows the ground truth vector field; the second and fourth columns are obtained using the proposed methodology and the one in [50], respectively; and the third and fifth columns are the corresponding errors by comparing them with the ground truth.

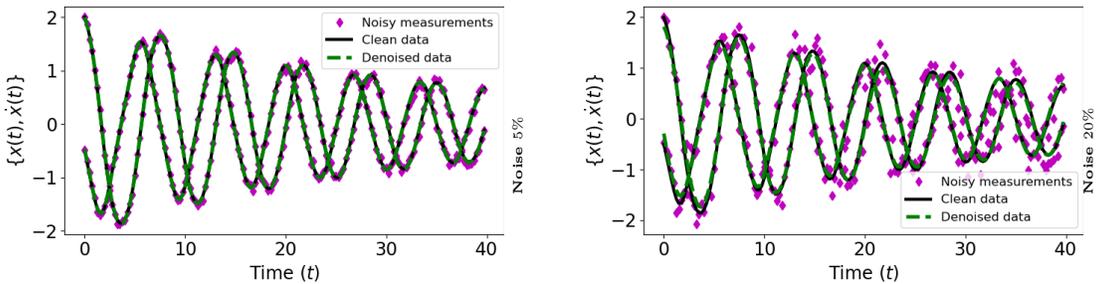


Figure 6.: Pendulum example: The figures show the ground truth, noisy measurements, and denoised data obtained from the implicit network.

interest in medical applications, where often data comes at quite irregular time intervals or when the sources of information are different. A similar problem has been discussed in [46], but for this, an initial condition is estimated using irregular trajectory points using ODE-RNN. It is then followed by integrating neural ODEs. Assessing the initial condition from the given irregular points is quite challenging, especially when the sequence is long, and also integrating a long sequence imposes additional challenges. Moreover, although the data can be collected at irregular intervals, all dependent variables still need to be measured/estimated on the same time grid. On the other hand, we can readily apply our proposed methods when all dependent variables are collected on different time-grids because of learning an implicit representation of the data.

We here present the framework for two-dimensional problems; however, it readily extends arbitrary dimensional dynamics. Let us consider a dynamical model as:

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t)), \tag{13}$$

where $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2] \in \mathbb{R}^2$. Next, assume that the variable \mathbf{x}_1 is measured at the time-grid $T_1 = \{t_1^{(1)}, \dots, t_n^{(1)}\}$, whereas the variable \mathbf{x}_2 is collected at the time-grid $T_2 = \{t_1^{(2)}, \dots, t_m^{(2)}\}$ with $T_1 \neq T_2$. To learn a model for the vector field representing the dynamics for \mathbf{x} using measurements at an irregular time-grid, we construct an implicit representation for \mathbf{x} so that both variables can be estimated on the same time-grid (let us denote it by $T = \{t_1, \dots, t_p\}$) but with a constraint using measurements. Assume the implicit network and neural ODE defining the vector field are denoted by \mathcal{N}_θ^I and $\mathcal{N}_\phi^{\text{Dyn}}$. To train the network, we define the following loss function:

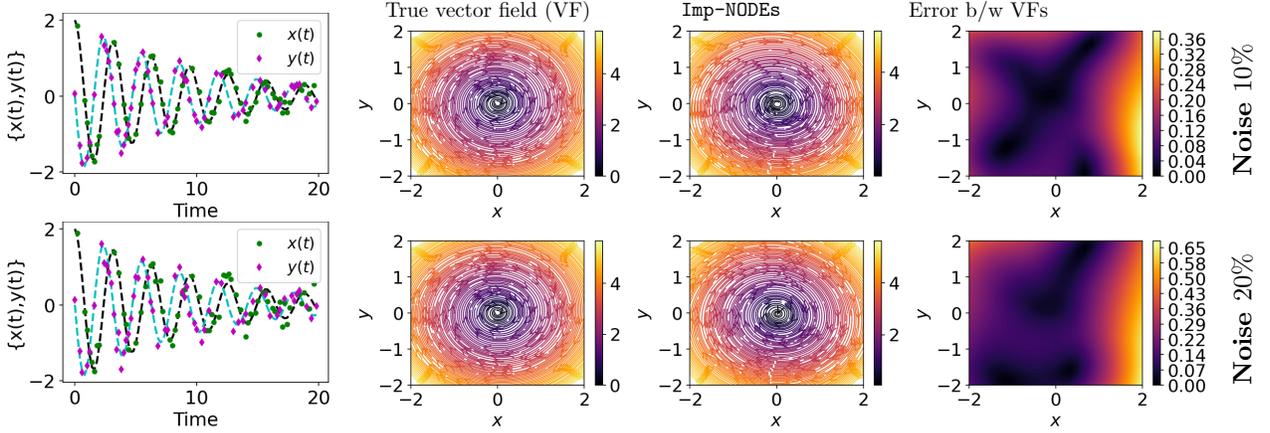


Figure 7.: Linear 2D example: The figures illustrate the ground truth and learned vector field using `Imp-NODEs`. The left-most plots show the irregularities in the sampled measurements, which clearly shows that the \mathbf{x} and \mathbf{y} do not share the same time grid. The dotted lines in the left-most figure show the ground truth full trajectories of \mathbf{x} and \mathbf{y} .

$$\lambda_{\text{MSE}} \left(\left\| \left[\mathcal{N}_{\theta}^1(t_i^{(1)}) \right]_1 - \mathbf{x}_1(t_i^{(1)}) \right\| + \left\| \left[\mathcal{N}_{\theta}^1(t_i^{(1)}) \right]_2 - \mathbf{x}_1(t_i^{(2)}) \right\| \right) + \lambda_{\text{Grad}} \left\| \mathcal{N}_{\phi}^{\text{Dyn}}(\mathcal{N}_{\theta}^1(t_i)) - \frac{d}{dt} \mathcal{N}_{\theta}^1(t_i) \right\| + \lambda_{\text{Integral}} \left\| (\mathcal{N}_{\theta}^1(t_j) - \mathcal{N}_{\theta}^1(t_i)) - \int_{t_i}^{t_j} \mathcal{N}_{\phi}^{\text{Dyn}}(\mathcal{N}_{\theta}^1(\tau)) d\tau \right\|,$$

where $[\cdot]_k$ denotes its k -element.

Numerical example: Linear 2D

We illustrate the considered scenario using a linear 2D example, given by

$$\begin{aligned} \dot{\mathbf{x}}(t) &= 0.1 \cdot \mathbf{x}(t) + 2.0 \cdot \mathbf{y}(t), \\ \dot{\mathbf{y}}(t) &= 2.0 \cdot \mathbf{x}(t) - 0.1 \cdot \mathbf{y}(t). \end{aligned}$$

We collect data using an initial condition $[\mathbf{x}, \mathbf{y}] = [2, 0]$ with a time interval $\mathbf{dt} = 0.2$ in the time interval $[0, 20]$. We randomly collect 60% independently for the first and second dependent variable, followed by corrupting them using Gaussian white noise of $\mu = \{10\%, 20\%\}$, as shown in the left-most column of Figure 7. Also, we take the time-grid for prediction of the output of the implicit network as the uniform grid with $\mathbf{dt} = 0.2$ for the time-interval $[0, 20]$. For learning models for the vector field, we set $\lambda_{\text{Integral}} = 1.0$, $\lambda_{\text{Grad}} = 10^{-2}$, and $\lambda_{\text{MSE}} = 1.0$ for $\mu = 10\%$ and $\lambda_{\text{MSE}} = 0.1$ for $\mu = 20\%$. We also do early-stopping for $\mu = 20\%$. For time integration, we consider the time span \mathbf{dt} .

We show the estimates of the learned vector fields using the proposed methodology and are compared with the ground truth in Figure 7, illustrating faithful capturing the dynamics. Moreover, we can also recover the clean signal faithfully without any prior information about the noise and any pre-processing of the data.

7. Discussion and Conclusion

This work has presented a new approach for learning dynamical models from highly noisy time-series data and obtaining denoise data. Our framework blends universal approximation capabilities of deep neural networks with neural ODEs. The proposed scheme involves two networks to learn (approximately) an implicit representation of the measurement data and of the vector field. These networks are combined by enforcing that an ODE can explain the dynamics of the output of the implicit network. We also discussed its extension to second-order neural ODEs to learn second-order dynamical models using corrupted data. Furthermore,

we have presented that the proposed approach can readily handle arbitrary sampled points in time. The dependent variables need not be collected at the same time-grid. It is because we first construct an implicit representation of the data that does not require data to be at a particular grid.

In the future, we focus on utilizing the encoder-decoder framework combined with an implicit network to learn latent ODEs to explain even richer dynamics of the measurement data. Moreover, when the data are high-dimensional (e.g., coming from partial-differential equations), applying neural ODEs becomes computationally intractable. However, it is known that the dynamics often lie in a low-dimensional manifold. Therefore, in our future work, we aim to utilize the concept of low-dimensional embedding to make learning computationally more efficient for high-dimensional data. Furthermore, it would be interesting to make use of expert knowledge and physical law to have physics-obey neural ODEs so that the generalizability and extrapolation capabilities of models can be further improved. Moreover, the performance of the proposed approach depends on the hyper-parameters that weigh different terms in the loss function. Therefore, we focus on developing an automatic mechanism to determine these parameters.

References

- [1] S. H. Rudy, J. N. Kutz, and S. L. Brunton, “Deep learning of dynamics and signal-noise decomposition with time-stepping constraints,” *J. Comput. Phys.*, vol. 396, pp. 483–506, 2019.
- [2] J.-N. Juang, *Applied System Identification*. Prentice-Hall, 1994.
- [3] L. Ljung, *System Identification – Theory for the User*. Upper Saddle River, NJ: Prentice-Hall, 2nd ed., 1999.
- [4] S. A. Billings, *Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains*. John Wiley & Sons, 2013.
- [5] B. Ho and R. E. Kálmán, “Effective construction of linear state-variable models from input/output functions,” *Automatisierungstechnik*, vol. 14, no. 1-12, pp. 545–548, 1966.
- [6] J.-N. Juang and R. S. Pappa, “An eigensystem realization algorithm for modal parameter identification and model reduction,” *J. Guidance, Control, and Dyn.*, vol. 8, no. 5, pp. 620–627, 1985.
- [7] R. W. Longman and J.-N. Juang, “Recursive form of the eigensystem realization algorithm for system identification,” *J. Guidance, Control, and Dyn.*, vol. 12, no. 5, pp. 647–652, 1989.
- [8] J.-N. Juang, M. Phan, L. G. Horta, and R. W. Longman, “Identification of observer/Kalman filter Markov parameters-theory and experiments,” *J. Guidance, Control, and Dyn.*, vol. 16, no. 2, pp. 320–329, 1993.
- [9] M. Phan, L. G. Horta, J.-N. Juang, and R. W. Longman, “Linear system identification via an asymptotically stable observer,” *J. Opt. Th. Appl.*, vol. 79, no. 1, pp. 59–86, 1993.
- [10] M. Phan, J.-N. Juang, and R. W. Longman, “Identification of linear multivariable systems by identification of observers with assigned real eigenvalues,” *J. Astronautical Sci.*, vol. 40, no. 2, 1992.
- [11] R. E. Kalman, “A new approach to linear filtering and prediction problems,” *J. Fluid Engrg.*, vol. 82, no. 1, 1960.
- [12] P. J. Schmid, “Dynamic mode decomposition of numerical and experimental data,” *J. Fluid Mech.*, vol. 656, pp. 5–28, 2010.
- [13] J. H. Tu, C. W. Rowley, D. M. Luchtenburg, S. L. Brunton, and J. N. Kutz, “On dynamic mode decomposition: Theory and applications,” *J. Comput. Dyn.*, vol. 1, no. 2, pp. 391–421, 2014.
- [14] I. G. Kevrekidis, C. W. Gear, J. M. Hyman, P. G. Kevrekidis, O. Runborg, C. Theodoropoulos, *et al.*, “Equation-free, coarse-grained multiscale computation: Enabling microscopic simulators to perform system-level analysis,” *Comm. Math. Sci.*, vol. 1, no. 4, pp. 715–762, 2003.

- [15] H. U. Voss, P. Kolodner, M. Abel, and J. Kurths, “Amplitude equations from spatiotemporal binary-fluid convection data,” *Physical Rev. Lett.*, vol. 83, no. 17, p. 3422, 1999.
- [16] H. Ye, R. J. Beamish, S. M. Glaser, S. C. Grant, C.-H. Hsieh, L. J. Richards, J. T. Schnute, and G. Sugihara, “Equation-free mechanistic ecosystem forecasting using empirical dynamic modeling,” *Proc. Nat. Acad. Sci. U.S.A.*, vol. 112, no. 13, pp. E1569–E1576, 2015.
- [17] M. D. Schmidt, R. R. Vallabhajosyula, J. W. Jenkins, J. E. Hood, A. S. Soni, J. P. Wikswo, and H. Lipson, “Automated refinement and inference of analytical models for metabolic networks,” *Phy. Biology*, vol. 8, no. 5, p. 055011, 2011.
- [18] B. C. Daniels and I. Nemenman, “Automated adaptive inference of phenomenological dynamical models,” *Nature Comm.*, vol. 6, no. 1, pp. 1–8, 2015.
- [19] B. C. Daniels and I. Nemenman, “Efficient inference of parsimonious phenomenological models of cellular dynamics using S-systems and alternating regression,” *PLoS One*, vol. 10, no. 3, p. e0119821, 2015.
- [20] J. Bongard and H. Lipson, “Automated reverse engineering of nonlinear dynamical systems,” *Proc. Nat. Acad. Sci. U.S.A.*, vol. 104, no. 24, pp. 9943–9948, 2007.
- [21] M. Schmidt and H. Lipson, “Distilling free-form natural laws from experimental data,” *Science*, vol. 324, no. 5923, pp. 81–85, 2009.
- [22] S. L. Brunton, J. L. Proctor, and J. N. Kutz, “Sparse identification of nonlinear dynamics with control (SINDYc),” *IFAC-PapersOnLine*, vol. 49, no. 18, pp. 710–715, 2016.
- [23] N. M. Mangan, S. L. Brunton, J. L. Proctor, and J. N. Kutz, “Inferring biological networks by sparse identification of nonlinear dynamics,” *IEEE Trans. Molecular, Biological and Multi-Scale Comm.*, vol. 2, no. 1, pp. 52–63, 2016.
- [24] G. Tran and R. Ward, “Exact recovery of chaotic systems from highly corrupted data,” *Multiscale Modeling & Simulation*, vol. 15, no. 3, pp. 1108–1129, 2017.
- [25] H. Schaeffer, G. Tran, R. Ward, and L. Zhang, “Extracting structured dynamical systems using sparse optimization with very few samples,” *Multiscale Modeling & Simulation*, vol. 18, no. 4, pp. 1435–1461, 2020.
- [26] N. M. Mangan, J. N. Kutz, S. L. Brunton, and J. L. Proctor, “Model selection for dynamical systems via sparse regression and information criteria,” *Proc. the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 473, no. 2204, p. 20170009, 2017.
- [27] P. Goyal and P. Benner, “Discovery of nonlinear dynamical systems using a Runge-Kutta inspired dictionary-based sparse regression approach,” *arXiv preprint arXiv:2105.04869*, 2021.
- [28] M. Raissi and G. E. Karniadakis, “Hidden physics models: Machine learning of nonlinear partial differential equations,” *J. Comput. Phys.*, vol. 357, pp. 125–141, 2018.
- [29] S. Chen, S. A. Billings, and P. Grant, “Non-linear system identification using neural networks,” *Intern. J. Control*, vol. 51, no. 6, pp. 1191–1214, 1990.
- [30] R. Rico-Martinez and I. G. Kevrekidis, “Continuous time modeling of nonlinear systems: A neural network-based approach,” in *IEEE Intern. Conf. on Neural Networks*, pp. 1522–1525, 1993.
- [31] R. Gonzalez-Garcia, R. Rico-Martinez, and I. Kevrekidis, “Identification of distributed parameter systems: A neural net based approach,” *Computers & Chemical Engrg.*, vol. 22, pp. S965–S968, 1998.
- [32] M. Milano and P. Koumoutsakos, “Neural network modeling for near wall turbulent flow,” *J. Comput. Phys.*, vol. 182, no. 1, pp. 1–26, 2002.
- [33] Z. Lu, B. R. Hunt, and E. Ott, “Attractor reconstruction by machine learning,” *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 28, no. 6, p. 061104, 2018.

- [34] S. Pan and K. Duraisamy, “Long-time predictive modeling of nonlinear dynamical systems using neural networks,” *Complexity*, vol. 2018, 2018.
- [35] J. Pathak, Z. Lu, B. R. Hunt, M. Girvan, and E. Ott, “Using machine learning to replicate chaotic attractors and calculate Lyapunov exponents from data,” *Chaos: An Interdisciplinary J. Nonlinear Sci.*, vol. 27, no. 12, p. 121102, 2017.
- [36] J. Pathak, A. Wikner, R. Fussell, S. Chandra, B. R. Hunt, M. Girvan, and E. Ott, “Hybrid forecasting of chaotic processes: Using machine learning in conjunction with a knowledge-based model,” *Chaos: An Interdisciplinary J. Nonlinear Sci.*, vol. 28, no. 4, p. 041101, 2018.
- [37] P. R. Vlachas, W. Byeon, Z. Y. Wan, T. P. Sapsis, and P. Koumoutsakos, “Data-driven forecasting of high-dimensional chaotic systems with long short-term memory networks,” *Proc. the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 474, no. 2213, p. 20170844, 2018.
- [38] B. Lusch, J. N. Kutz, and S. L. Brunton, “Deep learning for universal linear embeddings of nonlinear dynamics,” *Nature Comm.*, vol. 9, no. 1, pp. 1–10, 2018.
- [39] N. Takeishi, Y. Kawahara, and T. Yairi, “Learning Koopman invariant subspaces for dynamic mode decomposition,” in *Advances in Neural Information Processing Systems*, vol. 31, pp. 1130–1140, 2017.
- [40] E. Yeung, S. Kundu, and N. Hodas, “Learning deep neural network representations for Koopman operators of nonlinear dynamical systems,” in *American Control Conference (ACC)*, pp. 4832–4839, IEEE, 2019.
- [41] K. Champion, B. Lusch, J. N. Kutz, and S. L. Brunton, “Data-driven discovery of coordinates and governing equations,” *Proc. Nat. Acad. Sci. U.S.A.*, vol. 116, no. 45, pp. 22445–22451, 2019.
- [42] M. Raissi, P. Perdikaris, and G. E. Karniadakis, “Multistep neural networks for data-driven discovery of nonlinear dynamical systems,” *arXiv preprint arXiv:1801.01236*, 2018.
- [43] M. Raissi, P. Perdikaris, and G. E. Karniadakis, “Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations,” *J. Comput. Phys.*, vol. 378, pp. 686–707, 2019.
- [44] M. Raissi, A. Yazdani, and G. E. Karniadakis, “Hidden fluid mechanics: Learning velocity and pressure fields from flow visualizations,” *Science*, vol. 367, no. 6481, pp. 1026–1030, 2020.
- [45] R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, “Neural ordinary differential equations,” in *Advances Neural Inform. Processing Sys.*, vol. 32, pp. 6571–6583, 2018.
- [46] Y. Rubanova, R. T. Q. Chen, and D. K. Duvenaud, “Latent ordinary differential equations for irregularly-sampled time series,” in *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [47] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [48] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, “Pytorch: An imperative style, high-performance deep learning library,” *arXiv preprint arXiv:1912.01703*, 2019.
- [49] S. L. Brunton, J. L. Proctor, and J. N. Kutz, “Discovering governing equations from data by sparse identification of nonlinear dynamical systems,” *Proc. Nat. Acad. Sci. U.S.A.*, vol. 113, no. 15, pp. 3932–3937, 2016.
- [50] A. Norcliffe, C. Bodnar, B. Day, N. Simidjievski, and P. Liò, “On second order behaviour in augmented neural ODEs,” in *Advances in Neural Information Processing Systems*, vol. 34, 2020.
- [51] V. Sitzmann, J. Martel, A. Bergman, D. Lindell, and G. Wetzstein, “Implicit neural representations with periodic activation functions,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020.
- [52] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, “Fast and accurate deep network learning by exponential linear units (ELUs),” *arXiv preprint arXiv:1511.07289*, 2015.

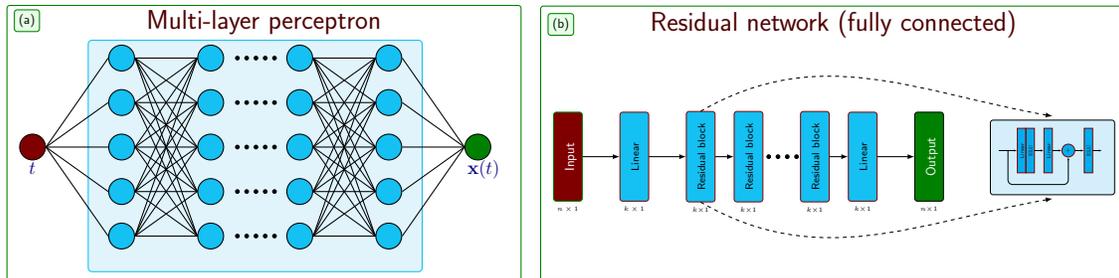


Figure 8.: The figure shows three potential simple architectures that can be used to learn either implicit representation or to approximate the underlying vector field. Diagram (a) is a simple multi-layer perceptron, and (b) is a residual-type network but fully connected.

Example	Networks	Neurons	Layers or residual blocks	Learning rates
Cubic oscillator	For implicit representation	20	4	10^{-3}
	For approximating vector field	20	4	10^{-3}
Pendulum	For implicit representation	32	4	$5 \cdot 10^{-4}$
	For approximating vector field	20	4	10^{-3}
Linear example	For implicit representation	20	3	$5 \cdot 10^{-4}$
	For approximating vector field	20	2	10^{-3}

Table 1.: The table shows the information about network architectures and learning rates.

A. Suitable Architectures and Chosen Hyper-parameter

Here, we briefly discuss neural network architectures suitable for our proposed approach. We require two neural networks for our framework, one for learning the implicit representation \mathcal{N}_θ^I and the second one $\mathcal{N}_\theta^{\text{Dyn}}$ is to learn the vector field. For implicit representation, we use a fully connected multi-layer perceptron (MLP) as depicted in Figure 8(a) with periodic activation functions (e.g., \sin) [51] which has shown its ability to capture finely detailed features as well as the gradients of a function. To approximate the vector field, we consider a simple residual-type network as illustrated in Figure 8(b) with *exponential linear unit* (ELU) as an activation function [52]. We choose ELU as the activation function since it is continuous and differentiable and resembles a widely used activation function, namely rectified linear unit (ReLU). Furthermore, to train implicit networks, we map the input data to $[-1, 1]$ as recommended in [51]. For the considered examples in the paper, we report the architecture designs (e.g., numbers of neural and layers) in Table 1.