

# THE DISCOVERY OF DYNAMICS VIA LINEAR MULTISTEP METHODS AND DEEP LEARNING: ERROR ESTIMATION

QIANG DU\*, YIQI GU†, HAIZHAO YANG‡, AND CHAO ZHOU§

**Abstract.** Identifying hidden dynamics from observed data is a significant and challenging task in a wide range of applications. Recently, the combination of linear multistep methods (LMMs) and deep learning has been successfully employed to discover dynamics, whereas a complete convergence analysis of this approach is still under development. In this work, we consider the deep network-based LMMs for the discovery of dynamics. We put forward error estimates for these methods using the approximation property of deep networks. It indicates, for certain families of LMMs, that the  $\ell^2$  grid error is bounded by the sum of  $O(h^p)$  and the network approximation error, where  $h$  is the time step size and  $p$  is the local truncation error order. Numerical results of several physically relevant examples are provided to demonstrate our theory.

**Key words.** Discovery of Dynamics; Convergence Analysis; Data-driven Modeling; Linear Multistep Methods; Deep Learning; Lorenz System.

**AMS subject classifications.** 65L06; 65L09; 65L20; 68T07;

**1. Introduction.** Dynamical systems are widely applied to characterize scientific principles and phenomena in various fields such as physics, biology, chemistry, economics, etc. In many situations, the observational data are accessible, whereas the underlying dynamics remain elusive. Data-driven discovery of dynamical systems is, therefore, an important research direction. There have been extensive study on data-driven discovery using Gaussian processes [22, 39, 40, 38], symbolic regression [45], S-systems formalism [11], sparse regression [6, 43, 57, 58], numerical PDE analysis [18], statistical learning [26], etc. Recently, along with the rapid advancements of deep learning, the discovery of dynamics using neural networks has also been proposed [37, 50, 44, 14, 15, 35, 49, 25, 53]. This paper studies high-order schemes for the discovery of dynamics using deep learning.

In numerical analysis, developing high-order methods is an important topic in many applications. In solving dynamical systems, high-order discretization techniques such as linear multistep methods (LMMs) and Runge-Kutta methods have been well-developed [4, 13, 30]. Recently, LMMs have been employed for the discovery of dynamics [41, 50, 54, 19]. In [19], a rigorous framework based on refined notions of consistency and stability is established to yield the convergence of LMM-based discovery for three popular LMM schemes (the Adams-Bashforth, Adams-Moulton, and Backwards Differentiation Formula schemes). However, the theory in [19] is specialized for methods that cannot provide a closed-form expression for the governing function, which is needed in many applications. Therefore, this paper studies the convergence theory of LMM schemes and deep learning, which can provide a closed-form description of the governing equation.

This paper concentrates on two types of discovery problems. The first type is to do the discovery on a trajectory of the dynamical system as in [41, 50, 54]. In this case, the observational data are collected from a specific trajectory, and the purpose is to identify the governing function on this trajectory with a closed-form expression in a form of a neural network, the parameters of

---

\*Department of Applied Physics and Applied Mathematics, and Data Science Institute, Columbia University, New York, NY (qd2125@columbia.edu)

†Department of Mathematics, National University of Singapore, 10 Lower Kent Ridge Road, Singapore, 119076 (matguy@nus.edu.sg)

‡Department of Mathematics, Purdue University, 150 N University St, West Lafayette, IN 47907, USA (haizhao@purdue.edu)

§Department of Mathematics, City University of Hong Kong, Kowloon Tong, Hong Kong. On leave from Department of Mathematics, National University of Singapore (chaozhou@cityu.edu.hk)

which are trained by minimizing the square residual of the corresponding LMM scheme. Through this work, we can forecast the future behavior of the same dynamics or predict the dynamics on nearby trajectories. The second type is the discovery on a compact region consisting of a bunch of trajectories on which the observational data are collected. The purpose is to identify the governing function in a connected compact region of the domain of the governing equation, which may not have been discussed in the literature.

In this paper, we perform a convergence analysis of these methods based on the LMM framework discussed in [19]. We first consider the LMMs using an abstract approximation set  $\mathcal{A}$ . The main result indicates that when using a  $p$ -th order LMM with a step size  $h$  in time, the  $\ell^2$  grid error of the obtained approximate function is bounded by  $O(\kappa_2(\mathbf{A}_h)(h^p + e_{\mathcal{A}}))$  where  $\kappa_2(\mathbf{A}_h)$  is the 2-condition number of the corresponding matrix of the LMM and  $e_{\mathcal{A}}$  is any upper bound of the approximation error between  $\mathcal{A}$  and the target function to be discovered (Theorem 5.1). Next, based on Theorem 5.1, we develop the error estimate of the network-based LMMs by using the recently developed approximation theory of deep networks [46, 47, 27, 48]. Note that Theorem 5.1 can also be used for the error estimate of LMMs using other approximation structures. Moreover, in connection with the stability theory developed in [19], we discuss the situations that  $\kappa_2(\mathbf{A}_h)$  is uniformly bounded with respect to  $h$ . Therefore, the  $\ell^2$  grid error decays to zero as  $h \rightarrow 0$  and the network size approaches to infinity.

In numerical examples, several artificial models and physically relevant benchmark problems are solved by the proposed network-based LMMs. It is observed the numerical error orders are consistent with our theory. We also conduct experiments to simulate the optimization errors in practice and the implicit regularization of deep learning. The results indicate that thanks to the implicit regularization, the network-based methods without auxiliary initial conditions can still find correct solutions numerically. Moreover, if the LMM scheme is unstable, the corresponding matrix is highly ill-conditioned, making traditional approximations (e.g., grid functions and polynomials) less robust, while the network approximation with gradient descent still manages to find solutions with similar ranges of errors as some of the stable counterparts.

This paper is organized as follows. In Section 2, background knowledge of dynamical systems and LMMs is introduced. In Section 3, we will introduce the LMM approach for the discovery of dynamics and discuss auxiliary conditions for unique recovery. In Section 4, the network-based LMM approach with ReLU neural networks is described. In Section 5, we discuss the convergence rate of the preceding approach with various LMM families. Numerical experiments are provided in Section 6 to validate the theoretical results. Finally, we conclude this paper in Section 7.

**2. Dynamical Systems.** In this section, we introduce some basic notations and definitions, as adopted by [19]. Most of the materials on LMMs can be found in [9, 10, 16].

**2.1. Initial Value Problem.** Suppose  $d > 0$  is the dimension of the dynamics, let us consider the following dynamical system with an initial condition

$$(2.1) \quad \frac{d}{dt}\mathbf{x}(t) = \mathbf{f}(\mathbf{x}(t)), \quad 0 < t < T,$$

$$(2.2) \quad \mathbf{x}(0) = \mathbf{x}_{\text{init}},$$

where  $\mathbf{x} \in C^\infty[0, T]^d$  is an unknown vector-valued state function;  $\mathbf{f} : \mathbb{R}^d \rightarrow \mathbb{R}^d$  is a given vector-valued governing function;  $\mathbf{x}_{\text{init}} \in \mathbb{R}^d$  is a given initial vector. To seek a numerical solution, one usually discretizes the problem by setting equidistant grid points in the time interval. Let  $N > 0$  be an integer, we define  $h := T/N$  and  $t_n = nh$  for  $n = 0, 1, \dots, N$ . The objective for solving the initial value problem in (2.1)-(2.2) is to find an approximate value  $\mathbf{x}_n \approx \mathbf{x}(t_n)$  for each  $n$  when  $\mathbf{f}(\mathbf{x})$  is given.

**2.2. Linear Multistep Method.** LMMs are widely utilized in solving dynamical systems. Suppose  $\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_{M-1}$  are given states, then  $\mathbf{x}_n$  for  $n = M, M+1, \dots, N$  can be computed by the following linear  $M$ -step scheme,

$$(2.3) \quad \sum_{m=0}^M \alpha_m \mathbf{x}_{n-m} = h \sum_{m=0}^M \beta_m \mathbf{f}(\mathbf{x}_{n-m}), \quad n = M, M+1, \dots, N,$$

where  $\alpha_m, \beta_m \in \mathbb{R}$  for  $m = 0, 1, \dots, M$  are specified coefficients and  $\alpha_0$  is always nonzero. By the scheme, all  $\mathbf{x}_n$  are evaluated iteratively from  $n = M$  to  $n = N$ . In each step,  $\mathbf{x}_{n-M}, \dots, \mathbf{x}_{n-1}$  are all given or computed previously such that  $\mathbf{x}_n$  can be computed by solving algebraic equations. If  $\beta_0 = 0$ , the scheme is called explicit since  $\mathbf{x}_n$  does not appear on the right hand side of (2.3) and  $\mathbf{x}_n$  can be computed directly by

$$(2.4) \quad \mathbf{x}_n = \frac{1}{\alpha_0} \sum_{m=1}^M (h\beta_m \mathbf{f}(\mathbf{x}_{n-m}) - \alpha_m \mathbf{x}_{n-m}).$$

Otherwise, the scheme is called implicit and it requires solving nonlinear equations for  $\mathbf{x}_n$ . The first value  $\mathbf{x}_0$  is simply set as  $\mathbf{x}_0 = \mathbf{x}_{\text{init}}$ , while other initial values  $\mathbf{x}_1, \dots, \mathbf{x}_{M-1}$  need to be computed by other approaches before performing the LMM if  $M > 1$ . Common types of LMMs include Adams-Bashforth (A-B) schemes, Adams-Moulton (A-M) schemes, and Backwards Differentiation Formula (BDF) schemes.

**2.3. Consistency.** An LMM is effective for a dynamical system only if it is consistent; that is, the discrete scheme (2.3) approximates the original differential equation (2.1) accurately as  $h$  is small enough. More specifically, we first define the local truncation error  $\tau_{h,n}$  as

$$(2.5) \quad \tau_{h,n} = \frac{1}{h} \sum_{m=0}^M \alpha_m \mathbf{x}(t_{n-m}) - \sum_{m=0}^M \beta_m \mathbf{f}(\mathbf{x}(t_{n-m})),$$

for  $n = M, M+1, \dots, N$ . Note that  $\tau_{h,n} \in \mathbb{R}^d$  is a numeric vector. It is clear that the local truncation error is defined by substituting the true function  $\mathbf{x}(t)$  into the discrete scheme (2.3), and measures the extent to which the true solution satisfies the discrete equation. By (2.1), we have the expression

$$(2.6) \quad \tau_{h,n} = \frac{1}{h} \sum_{m=0}^M \alpha_m \mathbf{x}(t_{n-m}) - \sum_{m=0}^M \beta_m \frac{d}{dt} \mathbf{x}(t_{n-m}).$$

Now we can define the notion of consistency. The LMM (2.3) is said to be consistent with the differential equation (2.1) if

$$(2.7) \quad \max_{M \leq n \leq N} \|\tau_{h,n}\|_{\infty} \rightarrow 0, \quad \text{as } h \rightarrow 0,$$

for any  $\mathbf{x} \in C^{\infty}[0, T]^d$ . Specifically, an LMM is said to have an order  $p$  if

$$(2.8) \quad \max_{M \leq n \leq N} \|\tau_{h,n}\|_{\infty} = O(h^p), \quad \text{as } h \rightarrow 0.$$

Method	$s$	$e(N)$	$t(N)$	$p$
$M$ -step A-B	0	$N - 1$	$N$	$M$
$M$ -step A-M	0	$N$	$N + 1$	$M + 1$
$M$ -step BDF	$M$	$N$	$N - M + 1$	$M$

Table 3.1: *The first involved index  $s$ , the last involved index  $e(N)$ , the total number of involved indices  $t(N)$ , and the truncation error order  $p$  for common types of LMMs.*

**3. Discovery of Dynamics.** In this section, we introduce the discovery of dynamics on a single trajectory, on which a time series of the state is available. Conventional LMMs with auxiliary conditions for this type of discovery are introduced. Note that these methods can be simply generalized for the discovery on a compact region, which will be discussed in Section 4.5.

The discovery of dynamics is essentially an inverse process of solving a dynamical system (2.1)-(2.2) with given observations on the state. That is, suppose that only the information of the state  $\mathbf{x}$  at the equidistant time steps  $\{t_n\}_{n=0}^N$  are provided, we would like to recover  $\mathbf{f}$ , namely, the governing function of the state.

**3.1. Linear Multistep Method.** Let  $\mathbf{x}(t) \in C^\infty([0, T])^d$  and  $\mathbf{f}(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^d$  be two vector-valued functions satisfying the dynamics (2.1), and we assume  $\mathbf{x}(t)$  and  $\mathbf{f}(\cdot)$  are both unknown. Now given  $\mathbf{x}_n = \mathbf{x}(t_n)$  for  $n = 0, \dots, N$ , the objective is to determine  $\mathbf{f}(\cdot)$ , i.e. to find a closed-form expression for  $\mathbf{f}(\cdot)$  or to evaluate  $\mathbf{f}(\mathbf{x}_n)$  for all  $n$ . One effective approach is to build a discrete relation between  $\mathbf{x}_n$  and  $\mathbf{f}_n \approx \mathbf{f}(\mathbf{x}_n)$  by LMMs [19], namely,

$$(3.1) \quad h \sum_{m=0}^M \beta_m \mathbf{f}_{n-m} = \sum_{m=0}^M \alpha_m \mathbf{x}_{n-m}, \quad n = M, M+1, \dots, N,$$

where  $\mathbf{f}_n \in \mathbb{R}^d$  is an approximation of  $\mathbf{f}(\mathbf{x}_n)$ . Note that (3.1) directly follows the LMM scheme (2.3). Different from (2.3) that evaluates  $\mathbf{x}_n$  given  $\mathbf{f}$ , (3.1) computes  $\mathbf{f}_n$  from the data  $\mathbf{x}_n$ . It indicates the dynamics discovery is actually an inverse process of solving the dynamical system [19]. Moreover, we note that the components of  $\mathbf{f}(\cdot)$  can be discovered independently. Thus, in the remainder of this paper, without loss of generality, we work with a scalar-valued system to simplify (3.1) using notation in a scalar form as the following general equation,

$$(3.2) \quad h \sum_{m=0}^M \beta_m f_{n-m} = \sum_{m=0}^M \alpha_m x_{n-m}, \quad n = M, M+1, \dots, N.$$

It is worth noting that  $f_n$  may not be involved in (3.2) for some indices  $n$  between 0 and  $N$ . For example, in A-B schemes,  $f_N$  does not appear in (3.2) since  $\beta_0 = 0$ . In general, given an LMM, we use  $s$  and  $e(N)$  to denote the first and last indices such that  $f_s$  and  $f_{e(N)}$  are involved in (3.2) with non-zero coefficients (correspondingly,  $\beta_{M-s}$  and  $\beta_{N-e(N)}$  are both nonzero). We also write  $t(N) := e(N) - s + 1$  as the total number of  $f_n$  involved in (3.2). We briefly list  $s$ ,  $e(N)$ ,  $t(N)$ , and the truncation error orders  $p$  of A-B, A-M, and BDF schemes in Table 3.1.

**3.2. Auxiliary Conditions.** For each linear  $M$ -step method, it is supposed to compute all unknowns  $\{f_n\}_{n=s}^{e(N)}$  by the linear relation (3.2). In the following, we will use the special notation  $\vec{\cdot}$  and bold fonts to denote column vectors of size  $O(N)$ , distinguishing them from other vectors or

vector functions. We write

$$(3.3) \quad \vec{f}_h := [f_s \quad f_{s+1} \quad \cdots \quad f_{e(N)}]^T \in \mathbb{R}^{t(N)},$$

$$(3.4) \quad \vec{q}_h := \frac{1}{h} \begin{bmatrix} \sum_{m=0}^M \alpha_m x_{M-m} & \sum_{m=0}^M \alpha_m x_{M+1-m} & \cdots & \sum_{m=0}^M \alpha_m x_{N-m} \end{bmatrix}^T \in \mathbb{R}^{N-M+1},$$

and

$$(3.5) \quad B_h := \begin{bmatrix} \beta_{M-s} & \beta_{M-s-1} & \cdots & \beta_{N-e(N)} \\ & \beta_{M-s} & \beta_{M-s-1} & \cdots & \beta_{N-e(N)} \\ & & \ddots & \ddots & \ddots & \ddots \\ & & & \beta_{M-s} & \beta_{M-s-1} & \cdots & \beta_{N-e(N)} \end{bmatrix} \in \mathbb{R}^{(N-M+1) \times t(N)},$$

then (3.2) leads to the following linear system,

$$(3.6) \quad B_h \vec{f}_h = \vec{q}_h.$$

However, the number of equations and unknowns may not be equal in (3.2). For A-B and A-M schemes, it is insufficient to determine  $\{f_n\}_{n=s}^{e(N)}$  by (3.2) since equations are less than unknowns. This is also implied by that the linear system (3.6) is underdetermined. For this issue, a natural solution is to provide  $N_a := t(N) - (N - M + 1)$  auxiliary linear conditions to make  $\{f_n\}_{n=s}^{e(N)}$  unique. For example, we can compute  $N_a$  certain unknown  $f_n$  directly by first-order (derivative) finite difference method (FDM) using related data. For consistency, the selected FDM should be of the same error order as the LMM. Assume the LMM has order  $p$ , one straightforward way is to compute the initial  $N_a$  unknowns by one-sided FDM of order  $p$ , i.e.,

$$(3.7) \quad f_n = \frac{1}{h} \sum_{m=0}^p \gamma_m x_{n+m}, \quad n = s, s+1, \dots, s+N_a-1,$$

where  $\gamma_m$  are the corresponding finite difference coefficients. Note that (3.7) has the error estimate

$$(3.8) \quad \max_{s \leq n \leq s+N_a-1} |f_n - f(x(t_n))| = O(h^p), \quad \text{as } h \rightarrow 0.$$

If we write

$$(3.9) \quad c_h := \frac{1}{h} \begin{bmatrix} \sum_{m=0}^p \gamma_m x_{s+m} & \sum_{m=0}^p \gamma_m x_{s+1+m} & \cdots & \sum_{m=0}^p \gamma_m x_{s+N_a-1+m} \end{bmatrix}^T \in \mathbb{R}^{N_a},$$

then combining the LMM (3.2) and the initial condition (3.7) leads to the following augmented linear system

$$(3.10) \quad A_h \vec{f}_h = \begin{bmatrix} c_h \\ \vec{q}_h \end{bmatrix},$$

where

$$(3.11) \quad A_h := \begin{bmatrix} C \\ B_h \end{bmatrix} \quad \text{and} \quad C := \begin{bmatrix} I_{N_a} & O \end{bmatrix}$$

with  $\mathbf{I}_{N_a}$  being the  $N_a \times N_a$  identity matrix and  $\mathbf{O}$  being the zero matrix of size  $N_a \times (t(N) - N_a)$ . Clearly, (3.10) has a unique solution since the coefficient matrix is lower triangular with nonzero diagonals. Moreover, if  $M \ll N$ , the linear system (3.10) is sparse.

In general, as pointed out in [19], we can formulate the auxiliary conditions in various ways, not just as discussed above. Different auxiliary conditions, such as initial and terminal conditions have different effects on the stability and the convergence of the method, see further discussions in [19]. An interesting question is whether the regularization effect provided by the neural network approximations could help mitigate these effects.

**4. Neural Network Approximation.** In this section, we first introduce the concept of fully connected neural networks (FNNs) and their approximation properties. Next, the network-based LMMs for the discovery on a trajectory will be presented together with the discussion on implicit regularization. In the end, we will generalize the methods for the discovery on a compact region.

**4.1. Preliminaries.** We introduce the fully connected neural network (FNN) which is widely used in deep learning. Mathematically speaking, given an activation function  $\sigma$ ,  $L \in \mathbb{N}^+$ , and  $W_\ell \in \mathbb{N}^+$  for  $\ell = 1, \dots, L$ , an FNN is the composition of  $L$  simple nonlinear functions, called hidden layer functions, in the following formulation:

$$(4.1) \quad \hat{\phi}(\mathbf{x}; \boldsymbol{\theta}) := \mathbf{a}^T \mathbf{h}_L \circ \mathbf{h}_{L-1} \circ \dots \circ \mathbf{h}_1(\mathbf{x}) \quad \text{for } \mathbf{x} \in \mathbb{R}^d,$$

where  $\mathbf{a} \in \mathbb{R}^{W_L}$ ;  $\mathbf{h}_\ell(\mathbf{x}_\ell) := \sigma(\mathbf{W}_\ell \mathbf{x}_\ell + \mathbf{b}_\ell)$  with  $\mathbf{W}_\ell \in \mathbb{R}^{W_\ell \times W_{\ell-1}}$  and  $\mathbf{b}_\ell \in \mathbb{R}^{W_\ell}$  for  $\ell = 1, \dots, L$ . With the abuse of notations,  $\sigma(\mathbf{x})$  means that  $\sigma$  is applied entry-wise to a vector  $\mathbf{x}$  to obtain another vector of the same size.  $W_\ell$  is the width of the  $\ell$ -th layer and  $L$  is the depth of the FNN.  $\boldsymbol{\theta} := \{\mathbf{a}, \mathbf{W}_\ell, \mathbf{b}_\ell : 1 \leq \ell \leq L\}$  is the set of all parameters in  $\hat{\phi}$  to determine the underlying neural network. Common types of activation functions include the rectified linear unit (ReLU)  $\max(0, x)$  and the sigmoid function  $(1 + e^{-x})^{-1}$ .

**4.2. Approximation Property.** Now let us introduce existing results on the approximation property of ReLU FNNs. Given a function  $g$  on a compact subset  $\mathcal{S}$  in  $\mathbb{R}^d$ , we can define the modulus of continuity by

$$(4.2) \quad \omega_g(\lambda) = \sup\{|g(\mathbf{x}) - g(\mathbf{y})| : \|\mathbf{x} - \mathbf{y}\|_2 \leq \lambda, \mathbf{x}, \mathbf{y} \in \mathcal{S}\}, \quad \text{for any } \lambda \geq 0,$$

where  $\|\mathbf{x}\|_2 := \sqrt{x_1^2 + x_2^2 + \dots + x_d^2}$  is the Euclidean norm of a vector in  $\mathbb{R}^d$ . Suppose  $\Lambda$  is any subset in  $\mathbb{R}^d$ , we define the  $C^r$  norm in  $\Lambda$ ,

$$(4.3) \quad \|g\|_{C^r(\Lambda)} := \max\left\{\|\partial^\alpha g\|_{L^\infty(\Lambda)} : \|\alpha\|_1 \leq r, \alpha \in \mathbb{N}^d\right\}.$$

Besides, we define

$$(4.4) \quad R_\Lambda := \inf\{r > 0 : \Lambda \in [-r, r]^d\},$$

Approximation properties of ReLU FNNs for continuous functions and smooth functions are indicated as follows.

**PROPOSITION 4.1.** *Given any  $J, K \in \mathbb{N}^+$  and a function  $g$  on a compact subset  $\mathcal{S}$  of  $\mathbb{R}^d$ ,*

1. *if  $g \in C(\mathcal{S})$ , there exists a ReLU FNN  $\hat{\phi}$  with width  $3^{d+3} \max\{d \lfloor J^{1/d} \rfloor, J+1\}$  and depth  $12K + 2d + 14$  such that*

$$(4.5) \quad |\hat{\phi}(\mathbf{x}) - g(\mathbf{x})| \leq 19\sqrt{d}\omega_g(2R_\mathcal{S}J^{-2/d}K^{-2/d}), \quad \text{for any } \mathbf{x} \in \mathcal{S};$$

2. if  $g \in C^r(\mathcal{S})$  with  $r \in \mathbb{N}^+$ , there exists a ReLU FNN  $\hat{\phi}$  with width  $17r^{d+1}3^d d(J+2)\log_2(8J)$  and depth  $18r^2(K+2)\log_2(4K) + 2d$  such that

$$(4.6) \quad |\hat{\phi}(\mathbf{x}) - g(\mathbf{x})| \leq 170R_{\mathcal{S}}(r+1)^d 8^r \|g\|_{C^r(\mathcal{S})} J^{-2r/d} K^{-2r/d}, \quad \text{for any } \mathbf{x} \in \mathcal{S},$$

The estimate (4.5) directly follows Theorem 4.3 in [46], and the estimate (4.6) can be derived from Theorem 1.1 in [27] by generalizing the results in the regular domain  $[0, 1]^d$  to a compact subset  $\mathcal{S}$ .

Note that the error bounds in (4.5) and (4.6) suffer the curse of dimensionality; namely, they exponentially depend on the dimension of the whole space  $\mathbb{R}^d$ . However, if we are only interested in the approximation on a low-dimensional manifold rather than a general compact subset in  $\mathbb{R}^d$ , the following stronger results can be adopted.

**PROPOSITION 4.2.** *Given  $J, K \in \mathbb{N}^+$ ,  $\varepsilon \in (0, 1)$ ,  $\delta \in (0, 1)$ . Let  $\mathcal{M} \subset \mathbb{R}^d$  be a compact  $d_{\mathcal{M}}$ -dimensional Riemannian submanifold having condition number  $\tau_{\mathcal{M}}^{-1}$ , volume  $V_{\mathcal{M}}$ , and geodesic covering regularity  $G_{\mathcal{M}}$ , and define the  $\varepsilon$ -neighborhood of  $\mathcal{M}$  as*

$$(4.7) \quad \mathcal{M}_{\varepsilon} := \{\mathbf{x} \in \mathbb{R}^d : \inf_{\mathbf{y} \in \mathcal{M}} \|\mathbf{x} - \mathbf{y}\|_2 \leq \varepsilon\}.$$

Suppose  $g$  is a function defined in  $\mathcal{M}_{\varepsilon}$ ,

1. if  $g \in C(\mathcal{M}_{\varepsilon})$ , there exists a ReLU FNN  $\hat{\phi}$  with width  $3^{d_{\delta}+3} \max\{d_{\delta} \lfloor J^{1/d_{\delta}} \rfloor, J+1\}$  and depth  $12K + 2d_{\delta} + 14$  such that

$$(4.8) \quad |\hat{\phi}(\mathbf{x}) - g(\mathbf{x})| \leq 2\omega_f \left( 4R_{\mathcal{M}}\varepsilon((1-\delta)^{-1}\sqrt{d/d_{\delta}} + 1) \right) \\ + 19\sqrt{d}\omega_g \left( 4R_{\mathcal{M}}(1-\delta)^{-1}\sqrt{d/d_{\delta}}J^{-2/d_{\delta}}K^{-2/d_{\delta}} \right), \quad \text{for any } \mathbf{x} \in \mathcal{M}_{\varepsilon};$$

2. if  $g \in C^r(\mathcal{M}_{\varepsilon})$  with  $r \in \mathbb{N}^+$ , there exists a ReLU FNN  $\hat{\phi}$  with width  $17r^{d_{\delta}+1}3^{d_{\delta}}d_{\delta}(J+2)\log_2(8J)$  and depth  $18r^2(K+2)\log_2(4K) + 2d_{\delta}$  such that

$$(4.9) \quad |\hat{\phi}(\mathbf{x}) - g(\mathbf{x})| \leq 8\|g\|_{C^r(\mathcal{M}_{\varepsilon})} R_{\mathcal{M}}\varepsilon((1-\delta)^{-1}\sqrt{d/d_{\delta}} + 1) \\ + 170R_{\mathcal{M}}(r+1)^{d_{\delta}} 8^r (1-\delta)^{-1} \|g\|_{C^r(\mathcal{M}_{\varepsilon})} J^{-2r/d_{\delta}} K^{-2r/d_{\delta}}, \quad \text{for any } \mathbf{x} \in \mathcal{M}_{\varepsilon},$$

where

$$(4.10) \quad d_{\delta} := O\left(\frac{d_{\mathcal{M}} \ln(dV_{\mathcal{M}}G_{\mathcal{M}}\tau_{\mathcal{M}}^{-1}/\delta)}{\delta^2}\right) = O\left(\frac{d_{\mathcal{M}} \ln(d/\delta)}{\delta^2}\right)$$

is an integer such that  $d_{\mathcal{M}} \leq d_{\delta} \leq d$ .

Equation (4.8) in Proposition 4.2 is an immediate result of Theorem 1.2 in [46] and Equation (4.9) can be derived from Theorem 1.1 in [27] and Theorem 4.4 in [46] similarly. In Proposition 4.2, both the error bounds and the ReLU FNN sizes depend on  $d_{\delta}$  instead of  $d$  so that the curse of dimensionality is lessened. Note that when  $\delta$  is closer to 1,  $d_{\delta}$  is closer to  $d_{\mathcal{M}}$ , then the approximation actually occurs in a reduced space with dimension close to  $d_{\mathcal{M}}$  instead of the whole space  $\mathbb{R}^d$ .

The approximation properties of other FNNs are also studied. For example, the properties of the Floor-ReLU FNN and a special three-hidden-layer FNN can be found in [47] and [48], respectively. Also, dimension-independent error bounds of FNNs for the target functions in Barron space are investigated in [5]. It is also interesting to apply these approximation theories to develop error estimates of dynamics discovery as future work.

**4.3. Network-based Methods for Discovery.** Let us review the discovery of dynamics on a single trajectory introduced in Section 3. Indeed, the discovery by conventional LMMs is simple to implement, and the solution can be found by merely solving a linear system. However, the governing function  $\mathbf{f}$  is only computed at prescribed equidistant time steps, and the relation between  $\mathbf{f}$  and the state  $\mathbf{x}$  is still unknown. One strategy to overcome this limitation is to approximate each component of  $\mathbf{f}$  by functions of specific structures such as neural networks, polynomials, splines, etc. The approximate functions can be determined through optimizations and will serve as closed-form expressions for  $\mathbf{f}$ . In real applications, once  $\mathbf{f}$  has been recovered with an explicit expression, the future behavior of the  $\mathbf{x}$  on the same trajectory can be forecasted via solving (2.1)-(2.2) with the given initial condition. On the other hand, the behavior of the  $\mathbf{x}$  on nearby trajectories can also be predicted via solving (2.1)-(2.2) with perturbed initial conditions.

Among all structures of approximations, it is popular to employ neural networks in the discovery problems. Especially, when  $d$  is moderately large, it is convenient to use neural networks to approximate the governing functions with high-dimensional inputs, which is usually intractable for other structures. Therefore we focus on the network-based methods in this paper. Note that the proposed methods can be easily generalized for other structures of approximations.

We consider the neural network approximation based on the LMM scheme (3.2). Generally, we use  $\mathcal{N}_{\hat{\mathcal{M}}}$  to denote the set of all neural networks with a specified architecture of a size set  $\hat{\mathcal{M}}$ . For example,  $\mathcal{N}_{\hat{\mathcal{M}}}$  can be the set of all FNNs with the fixed size  $\hat{\mathcal{M}} = \{L, W\}$ , where  $L$  is the depth and  $W$  is the width. The notation  $\hat{\mathcal{M}} \rightarrow \infty$  means that some of the numbers in  $\hat{\mathcal{M}}$  go to infinity.

Now we introduce a network  $\hat{f}_{\hat{\mathcal{M}}}(\mathbf{z}) \in \mathcal{N}_{\hat{\mathcal{M}}}$  to approximate  $f(\mathbf{z})$ , an arbitrary component of  $\mathbf{f}(\cdot)$ . The method of neural network approximation can be developed by replacing each  $f_n$  with  $\hat{f}_{\hat{\mathcal{M}}}(\mathbf{x}_n)$  in (3.2), namely,

$$(4.11) \quad h \sum_{m=0}^M \beta_m \hat{f}_{\hat{\mathcal{M}}}(\mathbf{x}_{n-m}) = \sum_{m=0}^M \alpha_m x_{n-m}, \quad n = M, M+1, \dots, N,$$

where  $\mathbf{x}_n$  for  $n = 0, \dots, N$  are given sample locations.

Unfortunately, if  $\hat{\mathcal{M}}$  is too small, the degree of freedom of  $\mathcal{N}_{\hat{\mathcal{M}}}$  will be less than the number of equations in (4.11) and, hence, there is no  $\hat{f}_{\hat{\mathcal{M}}} \in \mathcal{N}_{\hat{\mathcal{M}}}$  such that (4.11) is satisfied precisely. Even if  $\hat{\mathcal{M}}$  is large enough, it is usually intractable to solve (4.11) for  $\hat{f}_{\hat{\mathcal{M}}}$  directly because of the nonlinear parametrization of neural networks. Consequently, in practice, we seek  $\hat{f}_{\hat{\mathcal{M}}}$  by minimizing the residual of (4.11) under a machine learning framework. Namely, we aim to find  $\hat{f}_{\hat{\mathcal{M}}} \in \mathcal{N}_{\hat{\mathcal{M}}}$  such that

$$(4.12) \quad J_h(\hat{f}_{\hat{\mathcal{M}}}) = \min_{u \in \mathcal{N}_{\hat{\mathcal{M}}}} J_h(u),$$

where

$$(4.13) \quad J_h(u) := \frac{1}{N-M+1} \sum_{n=M}^N \left| \sum_{m=0}^M \beta_m u(\mathbf{x}_{n-m}) - \sum_{m=0}^M h^{-1} \alpha_m x_{n-m} \right|^2.$$

However, similar to the underdetermined linear system (3.6) that has infinitely many solutions, there exist infinitely many sets of real numbers  $\{y_n\}_{n=s}^{e(N)}$  such that  $J_h(u) = 0$  providing

$$(4.14) \quad u(\mathbf{x}_n) = y_n, \quad \forall n.$$

For each set  $\{y_n\}_{n=s}^{e(N)}$ , if the degree of freedom of  $\mathcal{N}_{\hat{\mathcal{M}}}$  is large enough, there is always some  $u \in \mathcal{N}_{\hat{\mathcal{M}}}$  such that (4.14) is satisfied due to overfitting. In this situation,  $u$  is a global minimizer of  $J_h$ .



Consequently,  $J_h$  admits infinitely many global minimizers, all of which lead to  $J_h = 0$  but take distinct values at  $\{\mathbf{x}_n\}_{n=s}^{e(N)}$ . It implies a minimizer of  $J_h$  might be totally different from the target governing function we aim to approximate.

To ensure the uniqueness of the minimizer in the function space at grid points, we introduce auxiliary conditions and build an augmented loss function based on (4.13). For example, the initial condition (3.7) on the solution network  $\hat{f}_{\hat{\mathcal{M}}}$  is enforced by solving

$$(4.15) \quad J_{a,h}(\hat{f}_{\hat{\mathcal{M}}}) = \min_{u \in \mathcal{N}_{\hat{\mathcal{M}}}} J_{a,h}(u),$$

where

$$(4.16) \quad J_{a,h}(u) := \frac{1}{t(N)} \left( \sum_{n=s}^{s+N_a-1} \left| u(\mathbf{x}_n) - \frac{1}{h} \sum_{m=0}^p \gamma_m x_{n+m} \right|^2 + \sum_{n=M}^N \left| \sum_{m=0}^M \beta_m u(\mathbf{x}_{n-m}) - \sum_{m=0}^M h^{-1} \alpha_m x_{n-m} \right|^2 \right).$$

The augmented optimization above guarantees that  $\hat{f}_{\hat{\mathcal{M}}}(\mathbf{x}_n) = \hat{g}_{\hat{\mathcal{M}}}(\mathbf{x}_n)$  for  $n = s, \dots, e(N)$  providing  $J_{a,h}(\hat{f}_{\hat{\mathcal{M}}}) = J_{a,h}(\hat{g}_{\hat{\mathcal{M}}}) = 0$ , for any  $\hat{f}_{\hat{\mathcal{M}}}, \hat{g}_{\hat{\mathcal{M}}} \in \mathcal{N}_{\hat{\mathcal{M}}}$ .

Indeed, two networks that are equal at grids  $\{\mathbf{x}_n\}_{n=s}^{e(N)}$  are not necessarily equal on the whole trajectory  $\{\mathbf{x}(t) : 0 \leq t \leq T\}$ . Fortunately, it is shown for regression problems and partial differential equation problems, deep learning can generalize well [20, 32, 31, 29]. This means the closeness of two networks at a dense set of training inputs can lead to their closeness at other nearby inputs. It can be inferred that  $\hat{f}_{\hat{\mathcal{M}}}(\mathbf{x}(t)) \approx \hat{g}_{\hat{\mathcal{M}}}(\mathbf{x}(t))$  for  $0 \leq t \leq T$  providing  $\hat{f}_{\hat{\mathcal{M}}}(\mathbf{x}_n) = \hat{g}_{\hat{\mathcal{M}}}(\mathbf{x}_n)$  for  $n = s, \dots, e(N)$  for any  $\hat{f}_{\hat{\mathcal{M}}}, \hat{g}_{\hat{\mathcal{M}}} \in \mathcal{N}_{\hat{\mathcal{M}}}$  as long as  $N$  is moderately large.

**4.4. Implicit Regularization.** We discuss the implicit regularization effect [33, 23] of gradient descent in practical deep learning. For regression problems, if we use over-parameterized FNNs with the standard random initialization, gradient descent can lead to global convergence with a linear convergence rate under certain conditions [17, 8, 12, 56]. Similar results also exist in the problems of solving partial differential equations [29]. Even though the global convergence could be established with over-parametrization, global minimizers are typically not unique. It is interesting to investigate what global minimizers would be identified by gradient descent and how the training process would reduce fitting errors. To answer these questions, it has been shown that, in regression problems, the training of FNN first captures low-frequency components of the target function and then starts to eliminate the high-frequency fitting error [55, 28]. Similar work about this spectral bias of deep learning is discussed in [7, 36]. In sum, all the above discussions show that neural networks trained by gradient descent in regression problems have an implicit bias towards smooth functions with low frequencies among all possible neural networks that perfectly fit training data.

Now let us consider the preceding network-based LMM optimizations. Note that the loss function (4.13) without auxiliary conditions and the loss function (4.16) with auxiliary conditions are formally close to the  $\ell^2$  loss in regression problems. Especially, for BDF schemes,  $\beta_0 = 1$  and  $\beta_2 = \beta_3 = \dots = 0$ , so the loss functions (4.13) and (4.16) are exactly the  $\ell^2$  loss. Hence, it is conjectured that the implicit regularization discussed above can also be applied to the LMM optimizations. Namely, the gradient descent tends to find a very smooth function among all global minimizers. Consequently, if the target governing function is also smooth enough, the gradient descent is expected to find a good approximation either through (4.12) without auxiliary conditions, or through (4.15) with auxiliary conditions. Numerical experiments in Section 6 will validate this fact.

**4.5. Discovery on a Compact Region.** The network-based formulation (4.15)-(4.16) is specific for the discovery on a single trajectory from which the data are collected. More generally, we can build similar formulations for the discovery on a connected compact region, from which a set of trajectories can be sampled. Through such a discovery, we can recover the whole vector field in this subset.

Suppose  $\mathbf{x}(t; \tilde{\mathbf{x}}_0)$  is the solution of (2.1) with initial value  $\tilde{\mathbf{x}}_0$ . Let  $\Gamma$  be a compact subset in  $\mathbb{R}^d$ , then  $\Omega := \{\mathbf{x}(t; \tilde{\mathbf{x}}) : 0 \leq t \leq T, \tilde{\mathbf{x}} \in \Gamma\} \subset \mathbb{R}^d$  is a compact region filled with all trajectories starting from  $\Gamma$  with time period  $0 \leq t \leq T$ . In practice, suppose we are given a dataset  $\{\mathbf{x}_{n,n'} = \mathbf{x}(t_n; \tilde{\mathbf{x}}_{n'})\}_{n=0, \dots, N; n'=1, \dots, N'}$ , where  $\{\tilde{\mathbf{x}}_{n'}\}_{n'=1, \dots, N'}$  is a set of points densely distributed in  $\Gamma$ , and suppose  $\Omega$  is densely covered by  $\{\mathbf{x}_{n,n'}\}$ . We aim to use neural networks to approximate the governing function in the whole subset  $\Omega$ .

Note that (4.16) is a loss function with respect to one trajectory. For multiple trajectories, we can build a similar loss function by summing up all individual loss functions with respect to each trajectory. Specifically, let  $\hat{f}_{\hat{\mathcal{M}}}$  be a network that approximates a certain component of the governing function, then we can determine  $\hat{f}_{\hat{\mathcal{M}}}$  by

$$(4.17) \quad J_{a,h,\text{multi}}(\hat{f}_{\hat{\mathcal{M}}}) = \min_{u \in \mathcal{N}_{\hat{\mathcal{M}}}} J_{a,h,\text{multi}}(u),$$

where

$$(4.18) \quad J_{a,h,\text{multi}}(u) := \frac{1}{N'} \sum_{n'=1}^{N'} J_{a,h,n'}(u)$$

and

$$(4.19) \quad J_{a,h,n'}(u) := \frac{1}{t(N)} \cdot \left( \sum_{n=s}^{s+N_a-1} \left| u(\mathbf{x}_{n,n'}) - \frac{1}{h} \sum_{m=0}^p \gamma_m x_{i+m,n'} \right|^2 + \sum_{n=M}^N \left| \sum_{m=0}^M \beta_m u(\mathbf{x}_{n-m,n'}) - \sum_{m=0}^M h^{-1} \alpha_m x_{n-m,n'} \right|^2 \right).$$

Similar to the discovery on a single trajectory, the optimization (4.17)-(4.19) for multiple trajectories will be also effective without auxiliary conditions due to the implicit regularization.

**5. Convergence Analysis.** In this section, we consider the convergence of the preceding network-based dynamics discovery using LMMs, namely, the convergence from the global minimizer of the optimization to the exact governing function  $f$  as  $\hat{\mathcal{M}} \rightarrow \infty$  and  $h \rightarrow 0$ . The optimization with auxiliary initial conditions is taken as a special case for analysis. For the optimization with other auxiliary conditions, a similar argument can be applied.

**5.1. Error Estimates on a Trajectory.** We consider the error estimation of the discovery on the specific trajectory  $\mathcal{T} := \{\mathbf{x}(t) : 0 \leq t \leq T\}$ . For least-square optimizations, people are usually interested in the  $\ell^2$ -type error estimation. Therefore, let us introduce the following  $\ell^2$  seminorm with a given  $h > 0$ ,

$$(5.1) \quad |g|_{2,h} := \left( \frac{1}{N+1} \sum_{n=0}^N |g(\mathbf{x}_n)|^2 \right)^{\frac{1}{2}}, \quad \forall g \in C(\mathcal{T}).$$

Note that  $|\cdot|_{2,h}$  is not a norm in  $C(\mathcal{T})$  since  $|g|_{2,h} = 0$  does not imply  $g = 0$  in  $C(\mathcal{T})$ . However,  $|\cdot|_{2,h}$  acts as a norm in the space of all grid functions merely defined on  $\{\mathbf{x}_n\}_{n=0}^N$  [19].

As discussed above, for a specific LMM, some states in  $\{\mathbf{x}_n\}_{n=0}^N$  may not be involved in the scheme. For a fair evaluation, it is proper to study the convergence at all involved states, namely,  $\{\mathbf{x}_n\}_{i=s}^{e(N)}$ . Therefore, we rewrite  $|\cdot|_{2,h}$  as the LMM-related seminorm

$$(5.2) \quad |g|_{2,h} = \left( \frac{1}{t(N)} \sum_{n=s}^{e(N)} |g(\mathbf{x}_n)|^2 \right)^{\frac{1}{2}}, \quad \forall g \in C(\mathcal{T}).$$

Without ambiguity, we use the notation  $|\cdot|_{2,h}$  for all LMMs afterwards. If we write  $\{g(\mathbf{x}_n)\}_{n=s}^{e(N)}$  in a vector form, namely,

$$(5.3) \quad \vec{g} := [g(\mathbf{x}_s) \quad g(\mathbf{x}_{s+1}) \quad \cdots \quad g(\mathbf{x}_{e(N)})]^T,$$

then it follows

$$(5.4) \quad |g|_{2,h} = \frac{1}{t(N)^{1/2}} \|\vec{g}\|_2,$$

where  $\|\cdot\|_2$  is the Euclidean norm of a column vector.

First, let us reformulate the optimization (4.15)-(4.16) with an abstract admissible set, say,

$$(5.5) \quad J_{a,h}(\hat{f}_{\mathcal{A},h}) = \min_{u \in \mathcal{A}} J_{a,h}(u),$$

where  $J_{a,h}(u)$  is defined in (4.16) and  $\mathcal{A}$  is a general nonempty set of functions. We aim to estimate the distance between  $\hat{f}_{\mathcal{A},h}$  and  $f$ .

For a given LMM, recall that  $\mathbf{B}_h$  defined in (3.5) is constructed by lining up the LMM coefficients into rows and  $\mathbf{A}_h$  is defined in (3.10). We denote the 2-condition number of  $\mathbf{A}_h$  by  $\kappa_2(\mathbf{A}_h) = \|\mathbf{A}_h\|_2 \|\mathbf{A}_h^{-1}\|_2$ . The estimation is described as follows.

**THEOREM 5.1.** *In the dynamical system (2.1), suppose  $\mathbf{x} \in C^\infty([0, T])^d$  and  $\mathbf{f}$  is defined in  $\mathcal{T}'$ , a small neighborhood of  $\mathcal{T}$ . Let  $f$  be an arbitrary component of  $\mathbf{f}$ . Also, let  $N > 0$  be an integer and  $h := T/N$ , then we have*

$$(5.6) \quad \left| \hat{f}_{\mathcal{A},h} - f \right|_{2,h} < C \kappa_2(\mathbf{A}_h) (h^p + e_{\mathcal{A}}),$$

where  $\hat{f}_{\mathcal{A},h} \in \mathcal{A}$  is a global minimizer of  $J_{a,h}$  defined by (4.16) corresponding to an LMM with order  $p$ ;  $e_{\mathcal{A}}$  is any real number such that  $e_{\mathcal{A}} > \inf_{u \in \mathcal{A}} \sup_{\mathbf{x} \in \mathcal{T}'} |u(\mathbf{x}) - f(\mathbf{x})|$ ;  $C$  is a constant independent of  $h$  and  $\mathcal{A}$ .

*Proof.* Given  $h > 0$ , similar to (2.5), we can define the component-wise local truncation error by

$$(5.7) \quad \tau_{h,n} := \frac{1}{h} \sum_{m=0}^M \alpha_m x_{n-m} - \sum_{m=0}^M \beta_m f(\mathbf{x}(t_{n-m})).$$

Then by denoting

$$(5.8) \quad \vec{\tau}_h := [\tau_{h,M} \quad \tau_{h,M+1} \quad \cdots \quad \tau_{h,N}]^T,$$

$$(5.9) \quad \vec{f} := [f(\mathbf{x}_s) \quad f(\mathbf{x}_{s+1}) \quad \cdots \quad f(\mathbf{x}_{e(N)})]^T,$$

we have

$$(5.10) \quad \vec{\tau}_h = \vec{q}_h - \mathbf{B}_h \vec{f}.$$

By the hypothesis that the LMM has order  $p$ , there exists some  $C_1 > 0$  independent of  $h$  such that

$$(5.11) \quad \|\vec{\tau}_h\|_2 \leq (N - M + 1)^{\frac{1}{2}} \|\vec{\tau}_h\|_\infty < C_1 (N - M + 1)^{\frac{1}{2}} h^p.$$

On the other hand, since  $e_{\mathcal{A}} > \inf_{u \in \mathcal{A}} \sup_{\mathbf{x} \in \mathcal{T}'} |u(\mathbf{x}) - f(\mathbf{x})|$ , there exists a function  $v \in \mathcal{A}$  such that

$$(5.12) \quad |v(\mathbf{x}) - f(\mathbf{x})| \leq e_{\mathcal{A}}$$

for all  $\mathbf{x} \in \mathcal{T}'$ .

Also, write  $\mathbf{e}_h = \mathbf{c}_h - [\mathbf{I}_{N_a} \ \mathbf{O}] \vec{f}$ , where  $\mathbf{c}_h$  is defined in (3.9). Then by (3.8), there exists some constant  $C_2$  independent of  $h$  such that

$$(5.13) \quad \|\mathbf{e}_h\|_2 \leq N_a^{\frac{1}{2}} \|\mathbf{e}_h\|_\infty < C_2 N_a^{\frac{1}{2}} h^p.$$

Moreover, write

$$(5.14) \quad \vec{f}_{\mathcal{A},h} = \begin{bmatrix} \hat{f}_{\mathcal{A},h}(\mathbf{x}_s) & \hat{f}_{\mathcal{A},h}(\mathbf{x}_{s+1}) & \cdots & \hat{f}_{\mathcal{A},h}(\mathbf{x}_{e(N)}) \end{bmatrix}^T,$$

$$(5.15) \quad \vec{v} := \begin{bmatrix} v(\mathbf{x}_s) & v(\mathbf{x}_{s+1}) & \cdots & v(\mathbf{x}_{e(N)}) \end{bmatrix}^T.$$

With these notations, by (3.10), we immediately have

$$(5.16) \quad J_{a,h}(\hat{f}_{\mathcal{A},h}) = \frac{1}{t(N)} \left\| \mathbf{A}_h \vec{f}_{\mathcal{A},h} - \begin{bmatrix} \mathbf{c}_h \\ \vec{q}_h \end{bmatrix} \right\|_2^2 = \frac{1}{t(N)} \left\| \mathbf{A}_h \left( \vec{f}_{\mathcal{A},h} - \vec{f} \right) - \begin{bmatrix} \mathbf{e}_h \\ \vec{\tau}_h \end{bmatrix} \right\|_2^2.$$

Since  $\hat{f}_{\mathcal{A},h} \in \mathcal{N}_{\hat{\mathcal{M}}}$  is a global minimizer of  $J_{a,h}$ , it satisfies  $J_{a,h}(\hat{f}_{\mathcal{A},h}) \leq J_{a,h}(v)$ , namely,

$$(5.17) \quad \frac{1}{t(N)} \left\| \mathbf{A}_h \left( \vec{f}_{\mathcal{A},h} - \vec{f} \right) - \begin{bmatrix} \mathbf{e}_h \\ \vec{\tau}_h \end{bmatrix} \right\|_2^2 \leq \frac{1}{t(N)} \left\| \mathbf{A}_h \left( \vec{v} - \vec{f} \right) - \begin{bmatrix} \mathbf{e}_h \\ \vec{\tau}_h \end{bmatrix} \right\|_2^2,$$

which implies

$$(5.18) \quad \left\| \mathbf{A}_h \left( \vec{f}_{\mathcal{A},h} - \vec{f} \right) \right\|_2 - \left\| \begin{bmatrix} \mathbf{e}_h \\ \vec{\tau}_h \end{bmatrix} \right\|_2 \leq \left\| \mathbf{A}_h \left( \vec{v} - \vec{f} \right) \right\|_2 + \left\| \begin{bmatrix} \mathbf{e}_h \\ \vec{\tau}_h \end{bmatrix} \right\|_2.$$

Therefore,

$$(5.19) \quad \left\| \vec{f}_{\mathcal{A},h} - \vec{f} \right\|_2 \leq \|\mathbf{A}_h^{-1}\|_2 \left\| \mathbf{A}_h \left( \vec{f}_{\mathcal{A},h} - \vec{f} \right) \right\|_2 \leq \|\mathbf{A}_h^{-1}\|_2 \left( \left\| \mathbf{A}_h \left( \vec{v} - \vec{f} \right) \right\|_2 + 2 \left\| \begin{bmatrix} \mathbf{e}_h \\ \vec{\tau}_h \end{bmatrix} \right\|_2 \right).$$

As a consequence, by (5.11), (5.12), (5.13), and (5.19), it follows that

$$\begin{aligned} \left| \hat{f}_{\mathcal{A},h} - f \right|_{2,h} &= t(N)^{-\frac{1}{2}} \left\| \vec{f}_{\mathcal{A},h} - \vec{f} \right\|_2 \\ &\leq t(N)^{-\frac{1}{2}} \left[ \|\mathbf{A}_h\|_2 \|\mathbf{A}_h^{-1}\|_2 \|\vec{v} - \vec{f}\|_2 + 2 \|\mathbf{A}_h^{-1}\|_2 (\|\mathbf{e}_h\|_2^2 + \|\vec{\tau}_h\|_2^2)^{\frac{1}{2}} \right] \\ &\leq t(N)^{-\frac{1}{2}} \left[ t(N)^{\frac{1}{2}} \|\mathbf{A}_h\|_2 \|\mathbf{A}_h^{-1}\|_2 \cdot e_{\mathcal{A}} + 2 \|\mathbf{A}_h^{-1}\|_2 (C_2^2 N_a + C_1^2 (N - M + 1))^{\frac{1}{2}} h^p \right] \\ &\leq C \kappa_2(\mathbf{A}_h) (h^p + e_{\mathcal{A}}). \end{aligned}$$

with  $C$  independent of  $h$  and  $\mathcal{A}$ , which completes the proof.  $\square$

The error estimate given in Theorem 5.1 is general for any types of the admissible set  $\mathcal{A}$ . Specifically, we propose the error estimate of the discovery using neural networks. Note that  $\mathcal{T}$  is a one-dimensional Riemannian submanifold, combining Theorem 5.1 and Proposition 4.2 directly leads to the following result.

**THEOREM 5.2.** *Under the notations and hypothesis of Theorem 5.1, for any  $J, K \in \mathbb{N}^+$  and  $\delta \in (0, 1)$ , it satisfies:*

1. *If  $f \in C(\mathcal{T}')$  and  $\mathcal{N}_{\hat{\mathcal{M}}}$  consists of all ReLU FNNs with width  $3^{d_\delta+3} \max\{d_\delta \lfloor J^{1/d_\delta} \rfloor, J+1\}$  and depth  $12K + 2d_\delta + 14$ ,*

$$(5.20) \quad \left| \hat{f}_{\hat{\mathcal{M}},h} - f \right|_{2,h} < C\kappa_2(\mathbf{A}_h) (h^p + e_{NN}(J, K))$$

with

$$(5.21) \quad e_{NN}(J, K) = \sqrt{d}\omega_f \left( 4R_{\mathcal{T}}(1-\delta)^{-1} \sqrt{d/d_\delta} J^{-2/d_\delta} K^{-2/d_\delta} \right);$$

2. *If  $f \in C^r(\mathcal{T}')$  with  $r \in \mathbb{N}^+$  and  $\mathcal{N}_{\hat{\mathcal{M}}}$  consists of all ReLU FNNs with width  $17r^{d_\delta+1} 3^{d_\delta} d_\delta (J+2) \log_2(8J)$  and depth  $18r^2(K+2) \log_2(4K) + 2d_\delta$ , then (5.20) still holds with*

$$(5.22) \quad e_{NN}(J, K) = R_{\mathcal{T}}(r+1)^{d_\delta} 8^r (1-\delta)^{-1} \|f\|_{C^r(\mathcal{T}')} J^{-2r/d_\delta} K^{-2r/d_\delta},$$

where  $d_\delta = O(\ln(d/\delta)/\delta^2)$  is an integer such that  $1 \leq d_\delta \leq d$ ;  $R_{\mathcal{T}}$  is defined by (4.4);  $\omega_f(\cdot)$  is defined by (4.2);  $\hat{f}_{\hat{\mathcal{M}},h} \in \mathcal{N}_{\hat{\mathcal{M}}}$  is a global minimizer of  $J_{a,h}$  defined by (4.16) corresponding to an LMM with order  $p$ ;  $C$  is a constant independent of  $h, J, K, d$  and  $d_\delta$ . In particular, if  $\kappa_2(\mathbf{A}_h)$  is uniformly bounded for all  $h > 0$ , then

$$(5.23) \quad \lim_{J,K \rightarrow \infty, h \rightarrow 0} \left| \hat{f}_{\hat{\mathcal{M}},h} - f \right|_{2,h} = 0.$$

**REMARK 5.1.** *If  $J$  and  $K$  are large enough, the error bound  $e_{NN}(J, K)$  will be overwhelmed by  $h^p$ . This means the LMM truncation error will dominate the network approximation error if the network size is large enough. In this situation,  $\left| \hat{f}_{\hat{\mathcal{M}},h} - f \right|_{2,h}$  will decay to zero with the rate  $O(h^p)$ . Namely, the convergence rate has the same order as the LMM scheme.*

Similarly, we can develop the  $\ell^2$  error estimate for the discovery on multiple trajectories (4.17)-(4.19). It suffices to use preceding results to obtain an error estimate for each trajectory and take the mean square of all these estimates. Specifically, we define

$$(5.24) \quad |g|_{2,h,\text{multi}} = \left( \frac{1}{t(N)N'} \sum_{n'=1}^{N'} \sum_{n=s}^{e(N)} |g(\mathbf{x}_{n,n'})|^2 \right)^{\frac{1}{2}}, \quad \forall g \in C(\Omega),$$

then under the hypothesis of Theorem 5.1, it satisfies

$$(5.25) \quad \left| \hat{f}_{\hat{\mathcal{M}},h} - f \right|_{2,h,\text{multi}} < C\kappa_2(\mathbf{A}_h) (h^p + e_{\mathcal{A}}),$$

where  $\hat{f}_{\hat{\mathcal{A}},h} \in \mathcal{A}$  is a global minimizer of  $J_{a,h}$  defined by (4.18)-(4.19) corresponding to an LMM with order  $p$ , and  $e_{\mathcal{A}}$  is any real number such that  $e_{\mathcal{A}} > \inf_{u \in \mathcal{A}} \sup_{\mathbf{x} \in \Omega} |u(\mathbf{x}) - f(\mathbf{x})|$ .

In particular, we can derive the error estimates for the discovery using ReLU FNNs if the governing function is either continuous or  $C^r$  smooth from Proposition 4.1. Similar arguments apply to other types of neural networks or other structures of approximations.

**5.2. Uniform Boundedness of  $\kappa_2(\mathbf{A}_h)$ .** Next, we discuss the estimation of  $\kappa_2(\mathbf{A}_h)$ . This is a special case, corresponding to the  $\|\cdot\|_{2,h}$  norm, of the discussion on the stability of LMM for dynamics discovery made in [19]. Here, for completeness, we provide an alternative approach to derive a conclusion that is the same as that shown in [19]. First, we introduce the following lemma ([3]),

LEMMA 5.3. *Given the following triangular Toeplitz band matrix*

$$(5.26) \quad \mathbf{T}_N = \begin{bmatrix} c_0 & & & & \\ \vdots & \ddots & & & \\ c_M & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & c_M & \cdots & c_0 \end{bmatrix} \in \mathbb{R}^{N \times N}$$

with  $c_0 \neq 0$ , we define the associated polynomial by

$$(5.27) \quad p(z) = \sum_{i=0}^M c_i z^{M-i}.$$

If all roots of  $p(z)$  have modulus smaller than 1, then  $\kappa_2(\mathbf{T}_N)$  is uniformly bounded, i.e.  $\kappa_2(\mathbf{T}_N) < C$  for some  $C$  independent of  $N$ .

Then we have the following theorem to determine the uniform boundedness of  $\kappa_2(\mathbf{A}_h)$ ,

THEOREM 5.4. *Let  $\mathbf{A}_h$  be the matrix defined by (3.11), and  $p_h(z)$  be the following polynomial*

$$(5.28) \quad p_h(z) = \sum_{i=N-e(N)}^{M-s} \beta_i z^{M-s-i}.$$

If all roots of  $p_h(z)$  have modulus smaller than 1, then  $\kappa_2(\mathbf{A}_h)$  is uniformly bounded with respect to  $N$ .

*Proof.* Rewrite  $\mathbf{A}_h$  as  $2 \times 2$  blocks

$$(5.29) \quad \mathbf{A}_h = \begin{bmatrix} \mathbf{I}_{N_a} & \mathbf{O} \\ \mathbf{B}_{h,1} & \mathbf{B}_{h,2} \end{bmatrix},$$

where

$$(5.30) \quad \mathbf{B}_{h,1} = \begin{bmatrix} \beta_{M-s} & \cdots & \beta_{N-e(N)+1} \\ & \ddots & \vdots \\ & & \beta_{M-s} \\ \cdots & & \end{bmatrix} \in \mathbb{R}^{(N-M+1) \times N_a},$$

and

$$(5.31) \quad \mathbf{B}_{h,2} = \begin{bmatrix} \beta_{N-e(N)} & & & \\ \vdots & \ddots & & \\ \beta_{M-s} & \ddots & \ddots & \\ & \ddots & \ddots & \ddots \\ & & \beta_{M-s} & \cdots & \beta_{N-e(N)} \end{bmatrix} \in \mathbb{R}^{(N-M+1) \times (N-M+1)}.$$

Clearly,  $\|\mathbf{B}_{h,1}\|_2$  only depends on the LMM scheme and independent of  $N$ . By Lemma 5.3, both  $\|\mathbf{B}_{h,2}\|_2$  and  $\|\mathbf{B}_{h,2}^{-1}\|_2$  are uniformly bounded with respect to  $N$ . Therefore, the proof directly follows

$$\begin{aligned}
 (5.32) \quad \|\mathbf{A}_h\|_2 &= \max_{\|\mathbf{x}\|_2=1} \|\mathbf{A}_h \mathbf{x}\|_2 = \max_{\|\mathbf{x}\|_2=1} \left\| \begin{bmatrix} \mathbf{I}_{N_a} & \mathbf{O} \\ \mathbf{B}_{h,1} & \mathbf{B}_{h,2} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \right\|_2 \\
 &= \max_{\|\mathbf{x}\|_2=1} \left( \|\mathbf{x}_1\|_2^2 + \|\mathbf{B}_{h,1}\mathbf{x}_1 + \mathbf{B}_{h,2}\mathbf{x}_2\|_2^2 \right)^{\frac{1}{2}} \leq \max_{\|\mathbf{x}\|_2=1} \left( \|\mathbf{x}_1\|_2^2 + (\|\mathbf{B}_{h,1}\|_2 \|\mathbf{x}_1\|_2 + \|\mathbf{B}_{h,2}\|_2 \|\mathbf{x}_2\|_2)^2 \right)^{\frac{1}{2}} \\
 &\leq \left( 1 + (\|\mathbf{B}_{h,1}\|_2 + \|\mathbf{B}_{h,2}\|_2)^2 \right)^{\frac{1}{2}},
 \end{aligned}$$

and

$$\begin{aligned}
 (5.33) \quad \|\mathbf{A}_h^{-1}\|_2 &= \max_{\|\mathbf{x}\|_2=1} \|\mathbf{A}_h^{-1} \mathbf{x}\|_2 = \max_{\|\mathbf{x}\|_2=1} \left\| \begin{bmatrix} \mathbf{I}_{N_a} & \mathbf{O} \\ -\mathbf{B}_{h,2}^{-1}\mathbf{B}_{h,1} & \mathbf{B}_{h,2}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \right\|_2 \\
 &= \max_{\|\mathbf{x}\|_2=1} \left( \|\mathbf{x}_1\|_2^2 + \|\mathbf{B}_{h,2}^{-1}\mathbf{B}_{h,1}\mathbf{x}_1 + \mathbf{B}_{h,2}^{-1}\mathbf{x}_2\|_2^2 \right)^{\frac{1}{2}} \\
 &\leq \max_{\|\mathbf{x}\|_2=1} \left( \|\mathbf{x}_1\|_2^2 + \|\mathbf{B}_{h,2}^{-1}\|_2^2 (\|\mathbf{x}_2\|_2 + \|\mathbf{B}_{h,1}\|_2 \|\mathbf{x}_1\|_2)^2 \right)^{\frac{1}{2}} \leq \left( 1 + \|\mathbf{B}_{h,2}^{-1}\|_2^2 (1 + \|\mathbf{B}_{h,1}\|_2)^2 \right)^{\frac{1}{2}}.
 \end{aligned}$$

□

**REMARK 5.2.** For BDF schemes,  $\beta_1 = \dots = \beta_M = 0$ , and the corresponding  $\mathbf{B}_{h,2}$  is a diagonal matrix with diagonals  $\beta_0$ . So  $\mathbf{A}_h$  is always uniformly bounded for each  $M \in \mathbb{N}$ . This means the network-based dynamics discovery with BDF schemes for all  $M \in \mathbb{N}$  is convergent in the sense of (5.23).

**REMARK 5.3.** For A-B schemes,  $\mathbf{B}_{h,2}$  is diagonal if  $M = 1$ . Also, it is verified for  $2 \leq M \leq 6$ , all the roots of the associated polynomial  $p_h(z)$  have modulus smaller than 1 ([19]). Hence, by Theorem 5.4,  $\mathbf{A}_h$  is uniformly bounded for  $1 \leq m \leq 6$ . This means the network-based dynamics discovery with A-B schemes for  $1 \leq M \leq 6$  is convergent in the sense of (5.23).

**REMARK 5.4.** For A-M schemes with  $M \geq 2$ , it was proven in [19] that all the roots of the associated polynomial  $p_h(z)$  have a modulus greater than 1. In these cases,  $\kappa_2(\mathbf{A}_h)$  increases exponentially with respect to  $N$ , and hence the error bounds in Theorem 5.2 also increases exponentially. This means we have no guarantee of their convergence in theory. In spite of this, it is still possible to obtain convergent solutions as  $h \rightarrow 0$  in practice (see Section 6.1.4 and Appendix A)

**REMARK 5.5.** Note that [19] considered stability under norms other than  $|\cdot|_{2,h}$  as well, which also allowed the discussion of convergence for A-B family for which there are roots on the unit disc. In particular, it was shown that A-M scheme is marginally stable for  $M = 1$ , (see the definition in [19]) but remains convergent. Actually, in this case,  $\kappa_2(\mathbf{A}_h)$  increases linearly with respect to  $N = T/h$ . If the network size is large enough such that the network approximation error is dominated by  $O(h^p)$ , the error bounds in Theorem 5.2 will be  $C \cdot \frac{T}{h} \cdot h^p = O(h)$  since  $p = 2$ . This means A-M scheme with  $M = 1$  is convergent with order 1. Moreover, Theorem 5.2 can be modified for norms other than  $|\cdot|_{2,h}$  and condition number other than  $\kappa_2$ , resulting in various error bounds with special orders.

**6. Numerical Experiments.** In this section, several examples are provided to show the performance of dynamics discovery via deep learning in practical computation. We aim to compute the errors of various LMMs, estimate the orders of accuracy and compare them with the theoretical ones.

In the first, second and third examples, we conduct the discovery on a single trajectory  $\mathcal{T}$  described in Section 4.3, in which we define the following relative  $\ell^2$  error

$$(6.1) \quad e_{\hat{f}} = \left( \frac{1}{d} \sum_{j=1}^d \frac{\sum_{n=s}^{e(N)} |\hat{f}_j(\mathbf{x}_n) - f_j(\mathbf{x}_n)|^2}{\sum_{n=s}^{e(N)} |f_j(\mathbf{x}_n)|^2} \right)^{\frac{1}{2}},$$

where  $f_n$  for  $n = 1, \dots, d$  are components of the original governing function  $\mathbf{f}$ , and  $\hat{f}_n$  is the network approximating  $f_n$ . Note that  $\{\mathbf{x}_n\}_{n=s}^{e(N)}$  are exactly the grid points involved in the loss function, the error defined by (6.1) is actually an empirical error. For the deep learning, we name (6.1) as the training error or grid error. On the other hand, we are also interested in the generalization performance of the network approximation. So we also define the relative  $\ell^2$  error at testing points as

$$(6.2) \quad \tilde{e}_{\hat{f}} = \left( \frac{1}{d} \sum_{j=1}^d \frac{\int_{\mathcal{T}} |\hat{f}_j - f_j|^2 ds}{\int_{\mathcal{T}} |f_j|^2 ds} \right)^{\frac{1}{2}} = \left( \frac{1}{d} \sum_{j=1}^d \frac{\int_0^T |\hat{f}_j(\mathbf{x}(t)) - f_j(\mathbf{x}(t))|^2 \cdot \|\mathbf{f}(\mathbf{x}(t))\|_2 dt}{\int_0^T |f_j(\mathbf{x}(t))|^2 \cdot \|\mathbf{f}(\mathbf{x}(t))\|_2 dt} \right)^{\frac{1}{2}},$$

where the integral over  $\mathcal{T}$  is evaluated by Gauss quadrature. For the deep learning, we name (6.2) as the testing error. Both (6.1) and (6.2) are taken as metrics for evaluation.

In the fourth example, we conduct the discovery on a compact region  $\Omega$  described in Section 4.5. Similarly, we define the following training error

$$(6.3) \quad e_{\hat{f}} = \left( \frac{1}{d} \sum_{j=1}^d \frac{\sum_{n'=1}^{N'} \sum_{n=s}^{e(N)} |\hat{f}_j(\mathbf{x}_{n,n'}) - f_j(\mathbf{x}_{n,n'})|^2}{\sum_{n'=1}^{N'} \sum_{n=s}^{e(N)} |f_j(\mathbf{x}_{n,n'})|^2} \right)^{\frac{1}{2}},$$

and testing error

$$(6.4) \quad \tilde{e}_{\hat{f}} = \left( \frac{1}{d} \sum_{j=1}^d \frac{\int_{\Omega} |\hat{f}_j - f_j|^2 d\mathbf{x}}{\int_{\Omega} |f_j|^2 d\mathbf{x}} \right)^{\frac{1}{2}},$$

where the integral over  $\Omega$  is evaluated by Monte Carlo method.

The overall setting in all experiments is summarized as follows.

- **Environment** The experiments are performed in Python 3.8 environment. We utilize PyTorch library for neural network implementation and CUDA 11.0 toolkit for GPU-based parallel computing. All examples are implemented on a desktop.
- **Optimizer and hyper-parameters** The network-based optimization is solved by *adam* subroutine from PyTorch library. This subroutine implements the Adam algorithm in [21]. For all examples, the number of epochs  $N_I$  is set as  $3 \times 10^4$ , and use batch gradient descent. The learning rate in the  $n$ -th epoch, denoted as  $\delta_n$ , is set to decay exponentially with linearly decreasing powers from  $10^{-2}$  to  $10^{-4}$ , namely,

$$\delta_n = 10^{-2-2n/N_I}.$$

- **Network setting** The FNN with ReLU activation is taken for approximation, whose weights and biases are initialized via uniform distributions  $\mathbf{W}_l, \mathbf{b}_l \sim U(-\sqrt{W_{l-1}}, \sqrt{W_{l-1}})$ .



- **Generation of data** In the first example, the state data are generated directly by the explicit expression. In the second and third examples, no expression for the state is available. Hence we generate the state data by solving the dynamical system via the solver `ode45` in Matlab with tiny tolerances (`RelTol` =  $10^{-13}$ , `AbsTol` =  $10^{-13}$ ).

In the numerical implementation, the overall error is not only affected by the LMM discretization error and the network approximation error, but also by the optimization performance. In neural network optimization, it is usually difficult to find global minimizers numerically due to non-convexity. There is no existing optimizer that can guarantee to identify a global minimizer to the best of our knowledge. The optimization error is the difference between the actually identified neural network and the neural network associated with an arbitrary global minimizer. Consequently, for LMMs with uniformly bounded  $\kappa_2(\mathbf{A})$ , the overall error between the numerical solution and the target governing function consists of the LMM discretization error  $O(h^p)$ , the network approximation error determined by the network size, and the optimization error. We will validate and quantify the optimization error in our tests later.

**6.1. Problem with Accurate Data.** Let us consider the following model problem

$$(6.5) \quad \begin{cases} \dot{x}_1 = x_2, \\ \dot{x}_2 = -x_1, \\ \dot{x}_3 = 1/x_2^2, \\ [x_1, x_2, x_3]_{t=0} = [0, 1, 0], \end{cases} \quad t \in [0, 1],$$

whose state can be explicitly given by  $x_1 = \sin(t)$ ,  $x_2 = \cos(t)$ ,  $x_3 = \tan(t)$ . Thanks to the explicit expressions, we can directly take the accurate time-series  $\{x_1(t_n), x_2(t_n), x_3(t_n)\}_{n=1}^N$  for the test, and no error is brought to the data. Under this setting, the error on numerical solutions are only caused by the method. In this experiment, we focus on the deep learning discovery with auxiliary initial conditions (4.15)-(4.16).

**6.1.1. Network Size Test.** Note that Theorem 5.2 implies  $e_{\hat{f}} \sim O(h^p)$  as  $h \rightarrow 0$ , as long as the network is sufficiently deep and wide. However, in practice, the desired depth and width are usually unknown. So we first perform the discovery with networks of various sizes to find a decent network that is both effective in approximation and cheap in computation. Specifically, we use depth  $L = 2, 3, \dots, 6$ , width  $W = 10, 20, \dots, 2560$ , and  $h = 10^{-3}$ . The BDF-6 scheme is employed in this test. Therefore, the local truncation error is up to  $O(h^6) = O(10^{-18})$ , which is smaller than machine precision. Consequently, numerical errors in this case are mainly caused by network approximation (i.e., the difference of the network associated with a global minimizer of (4.12) and (4.15) and the target function) and network optimization (i.e., the difference of the networks associated with a local minimizer and a global minimizer of (4.12) and (4.15)). In Figure 6.1,  $e_{\hat{f}}$  and  $\tilde{e}_{\hat{f}}$  versus  $W$  for various  $L$  are presented. It is observed that both errors decrease quickly as  $W$  increases. On the other hand, the network with  $L = 5$  and  $W = 2560$  obtains the minimal error. We can also observe that for  $L = 5$ , the error decay becomes very slow after  $W = 640$ . Consequently, we choose the network with  $L = 5$  and  $W = 640$  for all tests afterward, since the computation when  $W = 640$  is not expensive and the overall error cannot be improved significantly furthermore.

**6.1.2. Quantification of Optimization Errors.** A special test is conducted to estimate the optimization errors. First, we set up three ReLU FNNs with  $L = 5$  and  $W = 640$ , denoted as  $\hat{f}_1^*$ ,  $\hat{f}_2^*$ ,  $\hat{f}_3^*$ , and use them to fit the three components of the governing functions in (6.5), respectively. We use a standard least-square regression in this fitting. Next, we consider the dynamical system

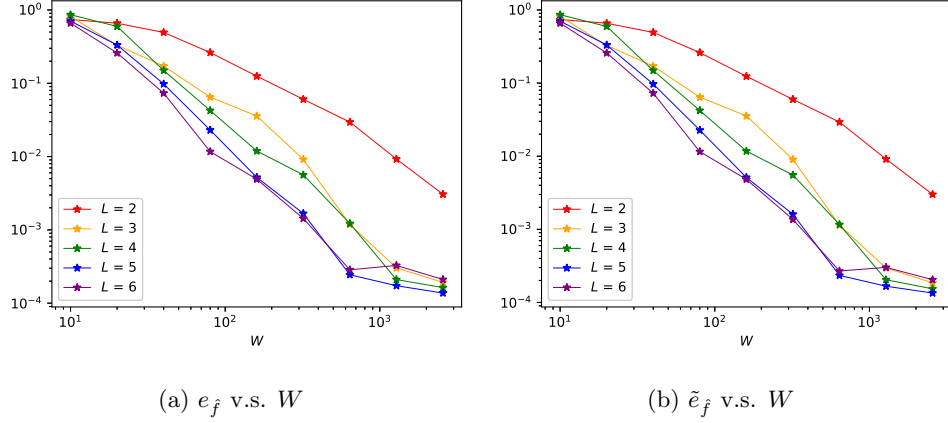


Fig. 6.1: Training error  $e_{\hat{f}}$  and testing error  $\tilde{e}_{\hat{f}}$  versus  $W$  of the model problem (6.5).

with the governing function being these FNNs, namely,

$$(6.6) \quad \begin{cases} \dot{x}_1 = \hat{f}_1^*, \\ \dot{x}_2 = \hat{f}_2^*, \\ \dot{x}_3 = \hat{f}_3^*, \\ [x_1, x_2, x_3]_{t=0} = [0, 1, 0], \end{cases} \quad t \in [0, 1].$$

We still use ReLU FNNs with  $L = 5$  and  $W = 640$  to do discovery on (6.6). Under this setting, the approximate networks have the same architecture as the target governing function, which implies the approximation error is automatically zero. Moreover, same as the preceding test, we take BDF-6 scheme with  $h = 10^{-3}$ , whose LMM discretization error is negligible. Therefore the obtained error should be dominated by the optimization error.

Finally, we obtain the training error  $e_{\hat{f}} = 3.451 \times 10^{-4}$  and testing error  $\tilde{e}_{\hat{f}} = 3.443 \times 10^{-4}$ , which reflects the optimization error caused by the current optimizer is around  $O(10^{-4})$ . This quantification indicates that there exists an error bottleneck around  $O(10^{-4})$  preventing the overall error from being reduced below it.

**6.1.3. Convergence Rate with Respect to  $h$ .** Next, we test the convergence rate of the deep learning discovery by varying  $h$  and using various LMM schemes. Recall the overall error consists of the LMM discretization error, the network approximation error, and the optimization error. To conduct appropriate tests on the convergence order in  $h$ , the network approximation error and the optimization error should be well controlled such that the LMM discretization error is the dominant error. For this purpose, we will conduct a series of tests to empirically identify a threshold  $h^* > 0$  such that the LMM discretization error is dominating the overall error when  $h > h^*$ . When  $h < h^*$ , although decreasing  $h$  would still reduce the overall error, it is difficult to observe the order of  $O(h^p)$  since, for example, the optimization error may be dominant.

Specifically, we assign  $h = 2^{-3}, \dots, 2^{-9}$ , fix the network width  $W = 640$ , and test A-B and BDF ( $M = 1, \dots, 4$ ) schemes, both of which are proved to have uniformly bounded matrices  $\mathbf{A}_h$ . The log-log error decay versus  $h$  for each scheme is presented in Figure 6.2. Recall the theoretical results in Section 5 imply that the training error of the  $M$ -step scheme should converge to zero with order  $M$ . According to Figure 6.2, there indeed exist some empirical threshold  $h^* > 0$  for

each scheme. It is shown in Figure 6.2 (a) that when  $h > h^*$ , deep learning-based LMMs can effectively discover the governing function on training sample points with error orders close to the theoretical ones. And it is shown in Figure 6.2 (b) that deep learning-based LMMs also have good generalization performance similar to the training error on sample points.

We would like to double-check that  $W = 640$  is an appropriate size and the approximation error is small enough for the convergence rate test with respect to  $h$ ; that is, the training errors are indeed dominated by  $O(h^p)$  when  $h > h^*$ , in which case the log-log error curves appear as straight line segments. For different  $M$  and  $h^*$ , we repeat the preceding test using width  $W = 1280$  and present the new training errors in Table 6.1. Table 6.1 shows that using  $W = 1280$  can not even get smaller errors in most cases, which excludes the possibility that the network approximation error is dominant. Note that decreasing  $h$  can reduce errors with an expected order as long as  $h > h^*$ , which excludes the possibility that the optimization error is dominant. Therefore, these numerical results show that the training errors are dominated by the LMM discretization when  $h > h^*$ , which is a suitable range of  $h$  for a convergence test.

**6.1.4. Convergence of A-M schemes.** Moreover, we perform a test using A-M schemes with  $1 \leq M \leq 4$ . Although no theoretical analysis is made on the convergence of A-M schemes with  $M \geq 2$  (see Remark 5.4), it is intriguing to investigate how the A-M schemes perform in practice. First, we conduct the network-based discovery with A-M schemes under the same framework as in Section 6.1.3. The training and testing errors versus  $h$  are shown in Figure 6.3. It is observed that both errors decrease as  $h$  decreases, though the errors decrease more slowly when  $h$  is smaller due to the optimization errors.

This result indicates that the network-based LMM with unstable schemes can still work effectively, obtaining solutions with small errors if  $h$  is small enough. However, comparative tests in Appendix A show that with unstable LMM schemes, using other approximations (e.g., grid functions and polynomials) are less robust, whose results are highly sensitive to the used solvers and their settings. This comparison implies that the network approximation is advantageous over other approximations in overcoming the ill-conditioning of the unstable schemes.

**6.2. Lorenz System.** Let us consider the 3-D Lorenz system which characterizes the chaotic dynamics for certain initial conditions and has a number of important applications including weather forecasting. The system is formulated as

$$(6.7) \quad \begin{cases} \dot{x}_1 = 10(x_2 - x_1), \\ \dot{x}_2 = x_1(28 - x_3) - x_2, \\ \dot{x}_3 = x_1x_2 - 8x_3/3, \end{cases} \quad t \in [0, T],$$

**6.2.1. Convergence Rate with Respect to  $h$ .** We continue testing the convergence rate with respect to  $h$  of the dynamics discovery via deep learning. As in the previous convergence test, the test is only valid when  $h$  is larger than a threshold  $h^*$  when the LMM discretization error is dominating the overall error. For simplicity, we only empirically choose  $W = 640$  since this width is large enough for the previous test. Specifically, we consider the long time behavior of the system (6.7) by setting  $T = 25$  and taking initial values  $[x_1, x_2, x_3]_{t=0} = [-8, 7, 27]$ . We assign  $h = 0.04, 0.02, \dots, 0.0025$  and take A-B ( $M = 1, \dots, 4$ ) and BDF ( $M = 1, \dots, 4$ ) schemes. First, we conduct the optimization with initial conditions (4.15)-(4.16). The error decay versus  $h$  is demonstrated in Figure 6.4. The dynamics of the true governing function and the approximate neural network obtained by A-B ( $M = 4, h = 0.0025$ ) are also presented in Figure 6.5, from which we observe that deep learning can identify the chaotic dynamics on training samples effectively.

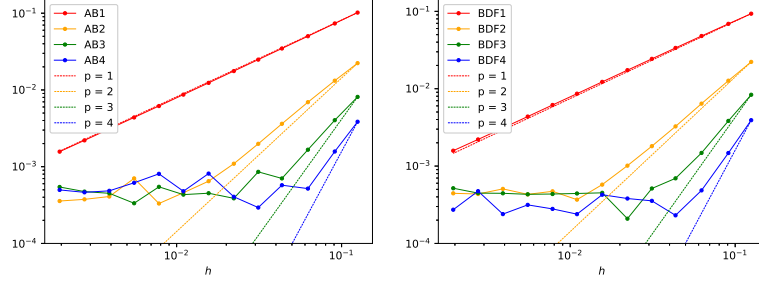
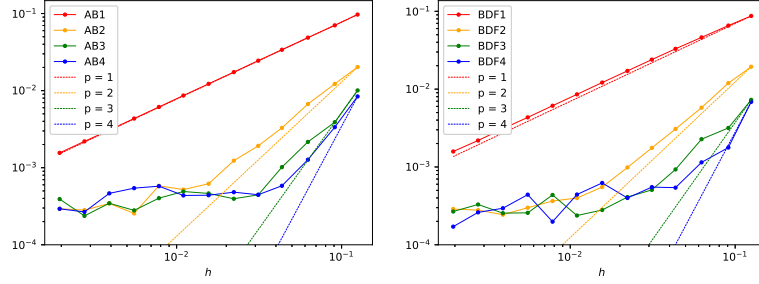
(a)  $e_{\hat{f}}$  v.s.  $h$ (b)  $\tilde{e}_{\hat{f}}$  v.s.  $h$ 

Fig. 6.2: Training error  $e_{\hat{f}}$  and testing error  $\tilde{e}_{\hat{f}}$  versus  $h$  via network-based A-B/BDF schemes of the model problem (6.5).

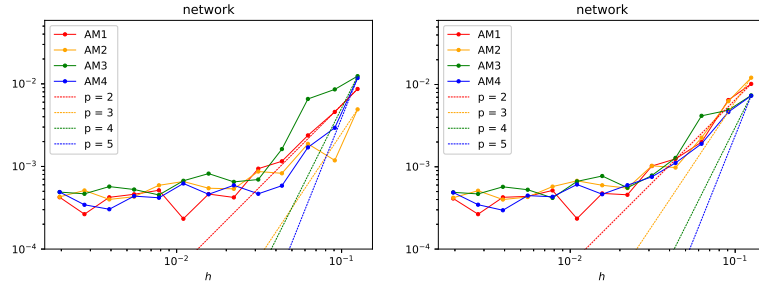
(a)  $e_{\hat{f}}$  v.s.  $h$ (b)  $\tilde{e}_{\hat{f}}$  v.s.  $h$ 

Fig. 6.3: Training error  $e_{\hat{f}}$  and testing error  $\tilde{e}_{\hat{f}}$  versus  $h$  via network-based A-M schemes of the model problem (6.5).

As discussed in Section 4.4, it is conjectured that auxiliary conditions may not be necessary to guarantee a correct solution because the implicit regularization has a bias towards the smoothest solution. To validate this fact, we conduct a comparative test by solving the optimization (4.12)-(4.13) with or without auxiliary conditions (ACs) in the loss function of the problem in (6.7). We take the same parameters as in the preceding test and visualize the error decay versus  $h$  for A-B schemes in Figure 6.4. We visualize the training error and loss versus training iterations in Figure 6.6. It is clear that when  $h = 0.02$ , the error of the A-B ( $M = 3$ ) scheme without auxiliary

Schemes		A-B			BDF		
$M, h^*$	$h$	$e_{\hat{f}, W=640}$	$e_{\hat{f}, W=1280}$	Diff	$e_{\hat{f}, W=640}$	$e_{\hat{f}, W=1280}$	Diff
$M = 1$ $h^* = 1/512$	1/8	1.019e-01	1.019e-01	6.755e-10	9.330e-02	9.987e-02	-6.576e-03
	1/32	2.485e-02	2.485e-02	-8.722e-07	2.431e-02	2.471e-02	-3.971e-04
	1/128	6.173e-03	6.179e-03	-5.386e-06	6.139e-03	6.142e-03	-3.145e-06
	1/512	1.561e-03	1.595e-03	-3.420e-05	1.582e-03	1.598e-03	-1.544e-05
$M = 2$ $h^* = 1/64$	1/8	2.234e-02	2.234e-02	-1.937e-12	2.222e-02	2.222e-02	1.128e-10
	1/16	6.928e-03	6.930e-03	-1.946e-06	6.405e-03	1.433e-02	-7.928e-03
	1/32	1.987e-03	2.045e-03	-5.865e-05	1.746e-03	1.877e-03	-1.304e-04
	1/64	6.110e-04	8.278e-04	-2.167e-04	5.380e-04	1.298e-03	-7.602e-04
$M = 3$ $h^* = 1/23$	1/8	8.117e-03	8.117e-03	-1.388e-17	8.354e-03	8.354e-03	1.105e-08
	1/11	4.035e-03	4.036e-03	-1.016e-06	3.836e-03	3.840e-03	-4.602e-06
	1/16	1.665e-03	1.744e-03	-7.959e-05	1.471e-03	1.950e-03	-4.786e-04
	1/23	6.735e-04	7.855e-04	-1.120e-04	6.531e-04	8.099e-03	-7.446e-03
$M = 4$ $h^* = 1/16$	1/8	3.852e-03	3.949e-03	-9.610e-05	3.920e-03	3.920e-03	1.996e-14
	1/11	1.577e-03	1.577e-03	-4.042e-09	1.472e-03	1.470e-03	<b>2.552e-06</b>
	1/16	5.179e-04	6.443e-04	-1.263e-04	5.413e-04	1.375e-03	-8.338e-04

Table 6.1: Training errors  $e_{\hat{f}}$  with  $W = 640$  and 1280 for various  $h$ 's using A-B and BDF schemes of the model problem (6.5).  $h^*$  denotes the step size threshold above which the error decreases as  $O(h^p)$  approximately. "Diff" denotes  $e_{\hat{f}, W=640} - e_{\hat{f}, W=1280}$ .

conditions is larger than the one with initial auxiliary conditions. The difference is more significant for A-B scheme ( $M = 4$ ) with  $h = 0.01$  and  $h = 0.02$ . However, when  $h$  is smaller, the errors of these two approaches seem to be almost the same. The comparison shows that, when  $h$  is relatively large, the approach with auxiliary conditions is more accurate, while both approaches with and without auxiliary conditions perform well when  $h$  is relatively small. Due to the non-uniqueness of networks approximately minimizing the loss function, networks with or without the auxiliary conditions can both reduce the loss functions well as shown by Figure 6.6 (c). However, reducing the loss function well does not imply the corresponding network converges to the right target function. When  $h$  is large, though the implicit regularization of deep learning can provide a smooth solution without auxiliary conditions, this solution may not be our target function and, hence, the error  $e_{\hat{f}}$  on the training grid points and the error  $\tilde{e}_{\hat{f}}$  on random grid points would be large as shown in Figure 6.6 (a) and (b) (left). When  $h$  is small, a larger number of training samples makes the loss function better restrict its local minimizers closer to the desired solution and, hence, both  $e_{\hat{f}}$  and  $\tilde{e}_{\hat{f}}$  becomes reasonable. The auxiliary conditions can better eliminate spurious local minimizers of the loss function and, hence, both  $e_{\hat{f}}$  and  $\tilde{e}_{\hat{f}}$  are reasonably small no matter  $h$  is large or small as shown in Figure 6.6 (a) and (b) (right).

**6.2.2. Prediction.** In real applications, we are interested in how well the discovered dynamics perform in making predictions. For this purpose, we first discover the system (6.7) with initial values  $[-8, 7, 27] + \varepsilon$  for  $\varepsilon = 0, 0.2$  and  $0.5$  by A-B scheme ( $M = 4, h = 0.0025$ ), obtaining networks  $\hat{\mathbf{f}}^0, \hat{\mathbf{f}}^{0.2}$  and  $\hat{\mathbf{f}}^{0.5}$ , respectively. Next, we solve the discovered system  $\dot{\mathbf{x}} = \hat{\mathbf{f}}^\varepsilon$  with initial values  $[-8, 7, 27] + \delta$  for  $\delta = 0, 0.2$  and  $0.5$  by the Matlab solver `ode45` with tiny tolerances (`RelTol` =  $10^{-13}$ , `AbsTol` =  $10^{-13}$ ), obtaining trajectories  $\hat{\mathbf{x}}^{\varepsilon, \delta}(t)$ . Moreover, we compute the trajectories of the exact system (6.7) with initial values  $[-8, 7, 27] + \delta$ , denoting as  $\mathbf{x}^\delta(t)$ , for comparison. The first component of these trajectories are presented in Figure 6.7.

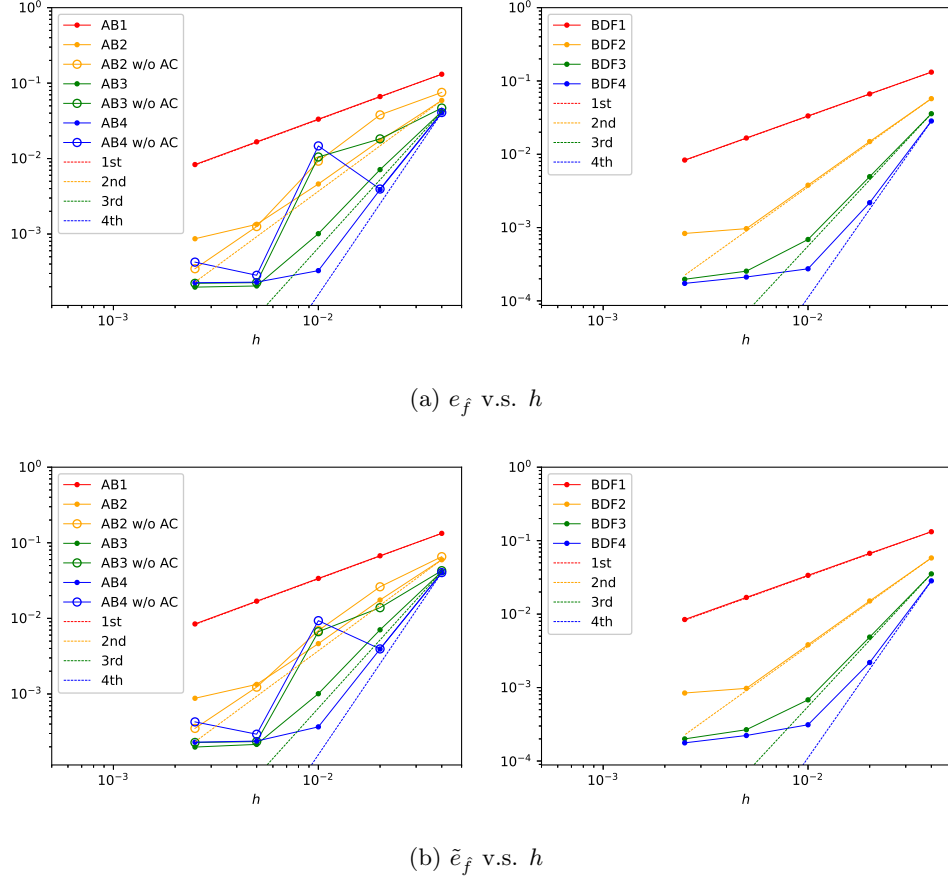


Fig. 6.4: Training error  $e_f$  and testing error  $\tilde{e}_f$  versus  $h$  via network-based A-B/BDF schemes with or without auxiliary conditions (ACs) of Lorenz system (6.7).

It can be observed that the predicted trajectories  $\hat{\mathbf{x}}^{\varepsilon, \delta}(t)$  become less accurate over time and ultimately fail to match the true trajectories after a particular time. The inaccuracy of the long-time prediction for the Lorenz system is imputed to its positive Lyapunov exponent, which results in exponential growth of any tiny initial error over time [52, 51]. Furthermore, the prediction performance also depends on the discrepancy between the initial value in prediction and the one for learning. Smaller discrepancy leads to better prediction. For example, in the case of  $\delta = 0$ , it is shown that  $\hat{\mathbf{x}}^{0,0}(t)$  moves consistently with  $\mathbf{x}^0(t)$  until  $t = 5.1$ , while  $\hat{\mathbf{x}}^{0,0.2}(t)$  and  $\hat{\mathbf{x}}^{0,0.5}(t)$  can only keep the consistency before  $t = 4.5$  and  $t = 1.5$ , respectively. Similarly, for  $\delta = 0.2$  and  $0.5$ , the trajectories  $\hat{\mathbf{x}}^{\delta, \delta}(t)$  has a longer accurately predicted period than  $\hat{\mathbf{x}}^{\varepsilon, \delta}(t)$  with  $\varepsilon \neq \delta$ . These numerical observations are due to the fact that only training samples of one trajectory are provided in deep learning and, hence, the recovered force term may not be accurate far away from the sampled trajectory. In real applications, historical data from a certain trajectory are given and the goal is to predict the future state in the same trajectory. Therefore, according to our numerical results, deep learning works well for this purpose.

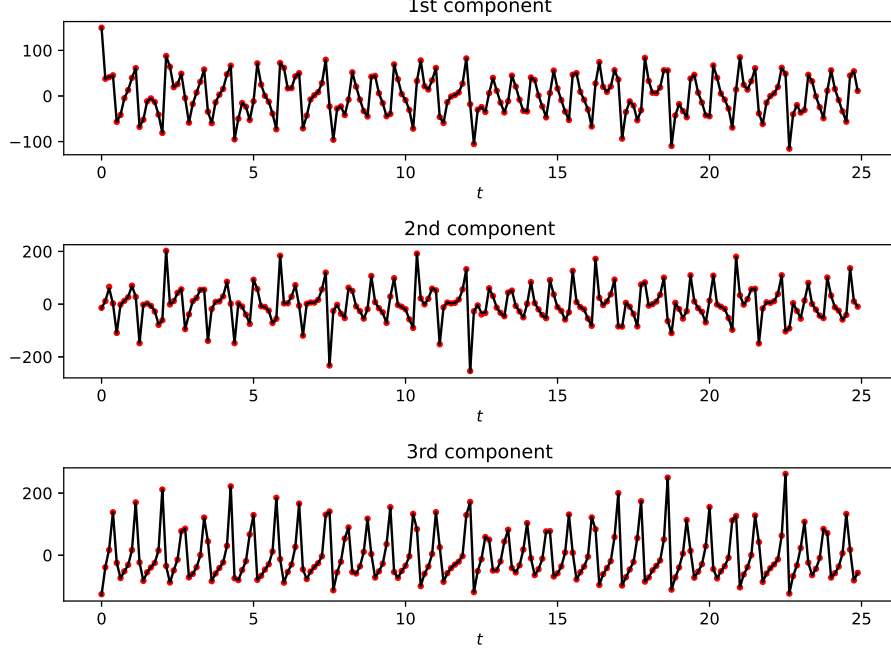


Fig. 6.5: The true governing function (solid black curves) and the approximate neural network (red circles) of Lorenz system (6.7).

**6.3. Glycolytic Oscillator.** We consider the model of oscillations in yeast glycolysis, which is a nonlinear biological system [11]. The model concentrates on 7 biochemical species:

$$(6.8) \quad \begin{cases} \dot{S}_1 = J_0 - \frac{k_1 S_1 S_6}{1 + (S_6/K_1)^q}, \\ \dot{S}_2 = 2 \frac{k_1 S_1 S_6}{1 + (S_6/K_1)^q} - k_2 S_2 (N - S_5) - k_6 S_2 S_5, \\ \dot{S}_3 = k_2 S_2 (N - S_5) - k_3 S_3 (A - S_6), \\ \dot{S}_4 = k_3 S_3 (A - S_6) - k_4 S_4 S_5 - \kappa (S_4 - S_7), \\ \dot{S}_5 = k_2 S_2 (N - S_5) - k_4 S_4 S_5 - k_6 S_2 S_5, \\ \dot{S}_6 = -2 \frac{k_1 S_1 S_6}{1 + (S_6/K_1)^q} + 2k_3 S_3 (A - S_6) - k_5 S_6, \\ \dot{S}_7 = \psi \kappa (S_4 - S_7) - k S_7, \end{cases} \quad t \in [0, T],$$

where the model parameters are taken from Table 1 in [11].

**6.3.1. Convergence Rate Test with Respect to  $h$ .** We continue testing the convergence rate with respect to  $h$  on the long time behavior of the system (6.8) with  $T = 10$  and the initial value  $[S_1, S_2, S_3, S_4, S_5, S_6, S_7]_{t=0} = \mathbf{S}_0$ , where

$$(6.9) \quad \mathbf{S}_0 = [1.125, 0.95, 0.075, 0.16, 0.265, 0.7, 0.092].$$

Similar to the preceding case, we assign  $h = 0.04, 0.02, \dots, 0.04/2^6$  and conduct the optimization (4.15)-(4.16) with A-B ( $M = 1, \dots, 4$ ) and BDF ( $M = 1, \dots, 4$ ) schemes. The error decay versus  $h$  is demonstrated in Figure 6.8. The dynamics of the true governing function and the neural network approximation obtained by A-B ( $M = 4, h = 0.04/2^6$ ) are presented in Figure 6.10. It is observed that when  $h$  is relatively large, the numerical convergence rates of all schemes are much

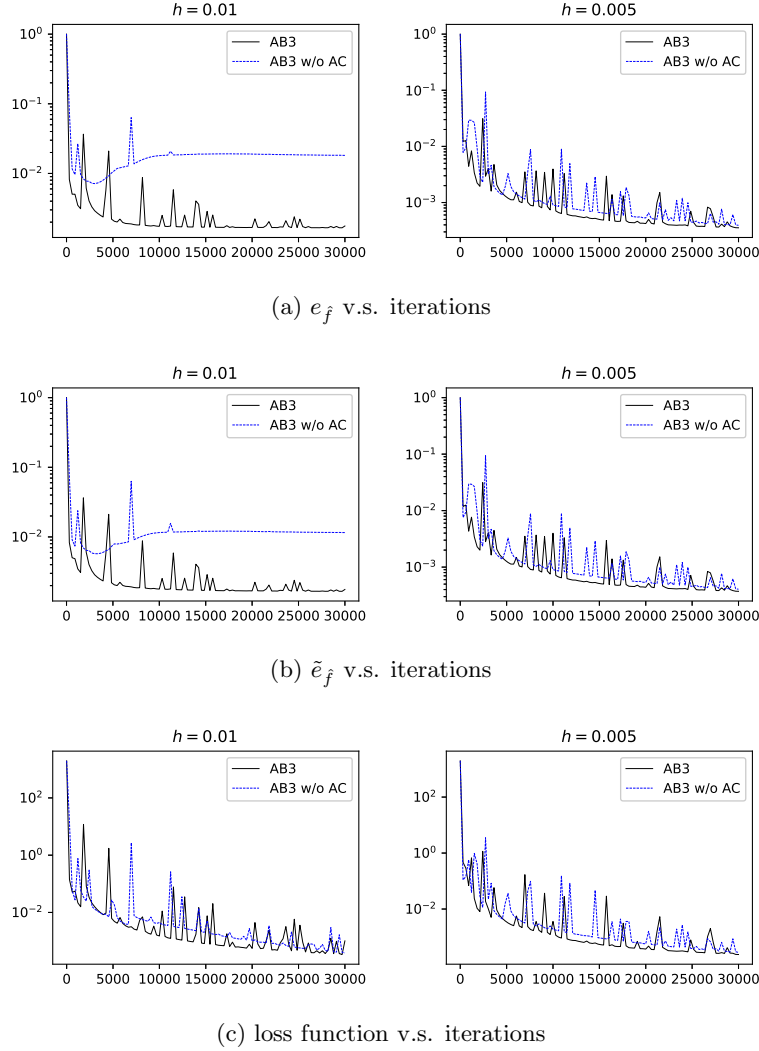


Fig. 6.6: The training error  $e_{\hat{f}}$ , testing error  $\tilde{e}_{\hat{f}}$  and loss function versus training iterations via network-based A-B schemes with or without auxiliary conditions (ACs) of Lorenz system (6.7).

lower than the theoretical ones. The same phenomenon can be observed if we recover the forcing term only on grid points by solving the linear system (3.10) without deep learning (See Figure 6.9). One explanation is that the low regularity of this system worsens the accuracy of LMMs. In Figure 6.10, it is clear that the governing function appears highly oscillatory with only  $C^0$  regularity. Even in this challenging case, high-order LMM schemes can still recover the governing function up to  $O(10^{-3})$  accuracy as  $h$  decreases.

**6.3.2. Prediction.** Similar to the preceding example, a prediction test is conducted for the glycolytic oscillator system. We compare the trajectories of the exact system (6.8) and the system discovered by the A-B scheme ( $M = 4$ ,  $h = 0.00125$ ) with training data generated with the initial value  $\mathbf{S}_0$ . The trajectories are computed with initial values  $\mathbf{S}_0 + \delta$  for  $\delta = 0, 0.05$  and  $0.2$ . In Figure 6.11, we present the first component of trajectories. The overall prediction performance in this example is better than that of the chaotic Lorenz system. The forecasted time-series when  $\delta = 0$  is very accurate. The forecasted time-series when  $\delta = 0.05$  and  $0.2$  are also reasonably



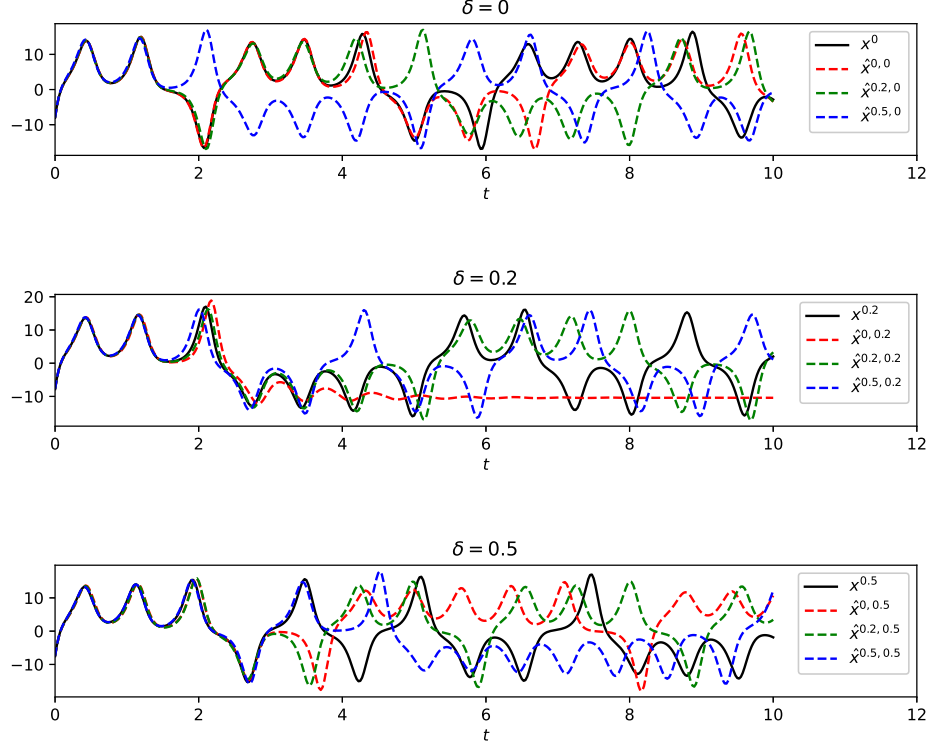


Fig. 6.7: The trajectories  $\mathbf{x}^\delta$  of the exact dynamics (6.7) (solid black curves) and the trajectories  $\hat{\mathbf{x}}^{\varepsilon, \delta}$  of the discovered dynamics (dashed red curves for  $\varepsilon = 0$ , dashed green curves for  $\varepsilon = 0.2$  and dashed blue curves for  $\varepsilon = 0.5$ ) with various initial values  $[-8, 7, 27] + \delta$ ,  $\delta = 0, 0.2, 0.5$ .

accurate, though the prediction error is obvious when the prediction time is large.

**6.4. Discovery on a Compact Region.** In this example, we consider the following model system

$$(6.10) \quad \begin{cases} \dot{x}_1 = 2x_1x_2, \\ \dot{x}_2 = x_1 + x_2, \\ [x_1, x_2]_{t=0} = \tilde{\mathbf{x}}, \end{cases} \quad t \in [0, 1].$$

The initial value point  $\tilde{\mathbf{x}}$  is chosen from the line segment  $\Gamma = \{(-0.5, x_2) : 0.5 \leq x_2 \leq 1\}$ . All the trajectories starting from  $\Gamma$  within  $t \in [0, 1]$  will form a compact region in  $\mathbb{R}^2$ , denoted as  $\Omega$ . Note that  $\Omega$  is enclosed with  $\Gamma$ ,  $\{(x_1(1; \tilde{\mathbf{x}}), x_2(1; \tilde{\mathbf{x}})) : \tilde{\mathbf{x}} \in \Gamma\}$  and two outside trajectories. We collect the data of discrete states in  $\Omega$ . Specifically, we choose  $N'$  points  $\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_{N'}$  by equidistantly partitioning  $\Gamma$  as the initial values. Next, we compute the trajectories  $\mathbf{x}(t; \tilde{\mathbf{x}}_{n'})$  for  $n' = 1, \dots, N'$  and take  $\{\mathbf{x}(t_n; \tilde{\mathbf{x}}_{n'})\}_{n=0, \dots, N; n'=1, \dots, N'}$  as the dataset. To display the data sampling clearly, we show the state points, trajectories and  $\Gamma$  for  $N = N' = 10$  in Figure 6.12, where the shaded region enclosed by  $\Gamma$  and outside trajectories is exactly  $\Omega$ .

**6.4.1. Convergence Rate with Respect to  $h$ .** Since the loss function of the discovery on a compact region is merely the sum of loss functions of the discovery on every involved trajectory

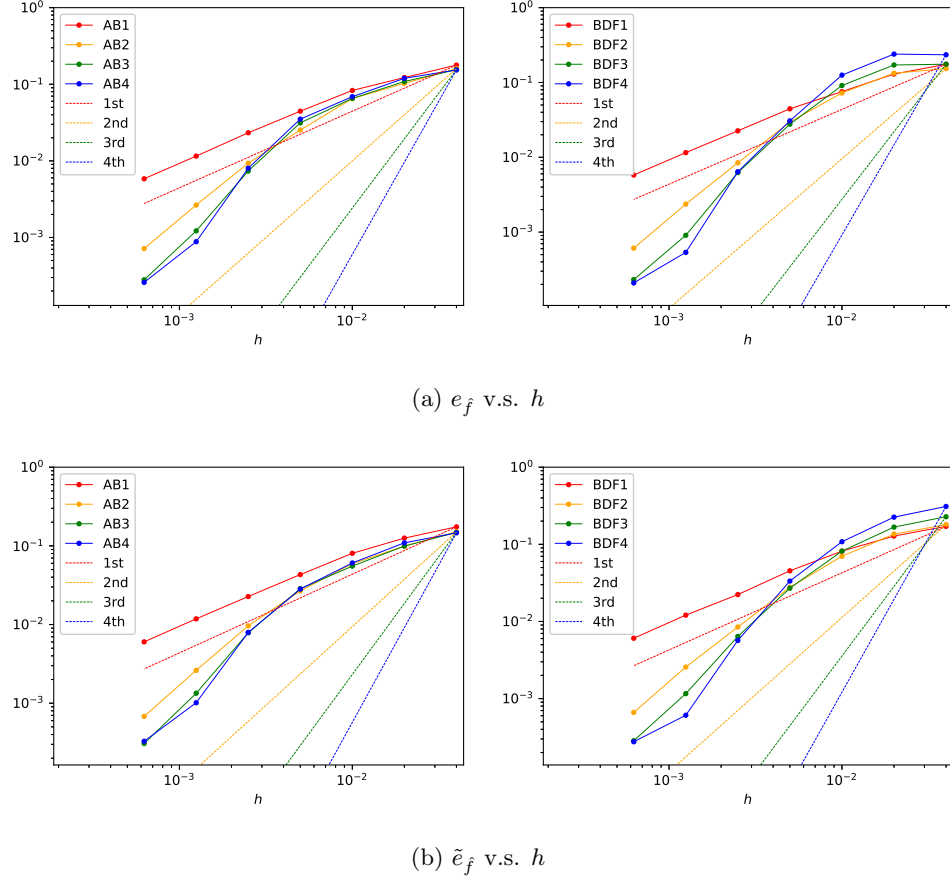


Fig. 6.8: Training error  $e_{\hat{f}}$  and testing error  $\tilde{e}_{\hat{f}}$  versus  $h$  via network-based A-B/BDF schemes of Glycolytic oscillator (6.8)

(see (4.18)-(4.19)), the implementation for the discovery on a compact region should share the same properties as the implementation on a trajectory, including the optimization errors and implicit regularization. The tests with respect to these properties will not be repeated in this example. Instead, we perform the test of the convergence rate with respect to  $h$  to validate the error estimate that  $e_{\hat{f}} = O(h^p)$  if the network size is large enough. We take A-B ( $M = 1, \dots, 4$ ) and BDF ( $M = 1, \dots, 4$ ) schemes for  $h = 0.1, 0.05, \dots, 0.1/2^4$ , then compute the training and testing errors (shown in Figure 6.13). The theoretical orders of error decay are observed when  $h$  is relatively large. While the overall error stops decreasing when  $h$  is too small due to the dominance of the optimization error. Specifically, the 2-D profiles of the obtained approximate networks, components of the true governing function and their errors are presented in Figure 6.14. The errors are observed to be below  $O(10^{-3})$  everywhere in  $\Omega$ .

**7. Conclusion.** This paper presents a rigorous convergence analysis of the network-based LMMs that discover unknown dynamical systems. The main result shows that the  $\ell^2$  grid error of the approximate function is bounded by  $O(\kappa_2(\mathbf{A}_h)(h^p + e_{\mathcal{A}}))$ , where  $\kappa_2(\mathbf{A}_h)$  is the 2-condition number of the corresponding matrix derived from the LMM scheme and  $e_{\mathcal{A}}$  is the approximation error of the admissible set. This result is combined with approximation properties of deep neural networks to develop the error estimates for network-based LMMs. We also characterize the root

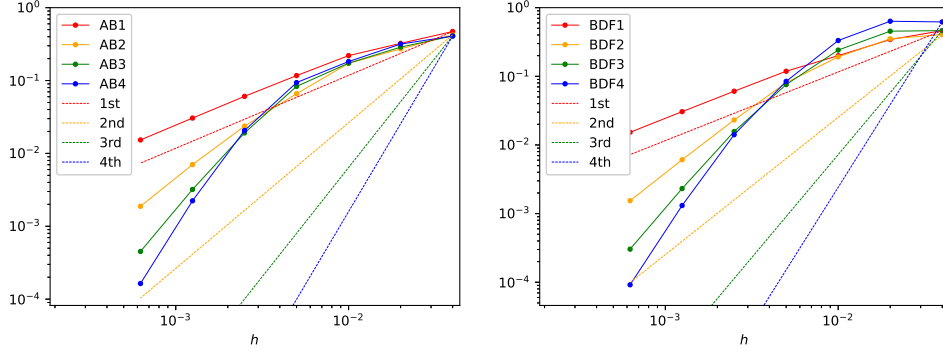


Fig. 6.9: Grid error  $e_f$  versus  $h$  via solving linear systems of *Glycolytic oscillator* (6.8).

condition to determine the uniform boundedness of  $\kappa_2(\mathbf{A}_h)$ . Besides, several numerical experiments are conducted to validate our theory. We observe that the error decaying orders of various LMMs are close to the theoretical ones.

In the experiments, we also test the network-based method either using formulations without auxiliary conditions or using unstable LMM schemes. In theory, we can not guarantee the uniqueness of the solution at grid points in the former case, and we do not have upper bounds for the discovery error in the latter case. However, in practice, deep learning with gradient descent can still find solutions with errors in the similar ranges of their stable counterparts. More traditional approximations, such as grid functions and polynomials, are less robust and sensitive to the choice of solvers in comparison (see Appendix A).

One limitation of our work is that the error estimation only quantifies the grid error, which is evaluated at the given sample locations. The generalization error out of sample locations is still theoretically unknown, though we observe excellent generalization performance in numerical experiments. Inspired by the works on generalization performance of deep learning for regression problems [20, 32, 31], decision problems [42] and PDEs [29], it is interesting to improve the error estimation from sample grid points to the whole trajectory. For example, the overlearning performance is studied in [42] using Rademacher complexity. Moreover, recurrent neural networks (RNNs) have been widely employed to build machine learning models of temporal data. The research on RNN generalization [1, 2, 34, 24] may shed light on the convergence analysis of the dynamics discovery.

Furthermore, our error analysis concentrates on the formulation with auxiliary conditions, while numerical tests show that the deep learning approach without auxiliary conditions can still perform well when the time step size is small enough. This might be due to the implicit regularization of the gradient descent and neural networks. Consequently, further investigation of the implicit regularization without auxiliary conditions is very interesting.

**Acknowledgments.** Q. D. is supported in part by the US NSF CCF-1704833 and DMS-2012562. Y. G. is supported by Singapore MOE AcRF Grants R-146-000-271-112. C. Z. is supported by Singapore MOE (Ministry of Educations) AcRF Grants R-146- 000-271-112 and R-146-000-284-114 as well as NSFC Grant No. 11871364. H. Y. was partially supported by the US National Science Foundation under award DMS-1945029.

## REFERENCES

- [1] Z. Allen-Zhu and Y. Li. Can sgd learn recurrent neural networks with provable generalization? In *Advances in*

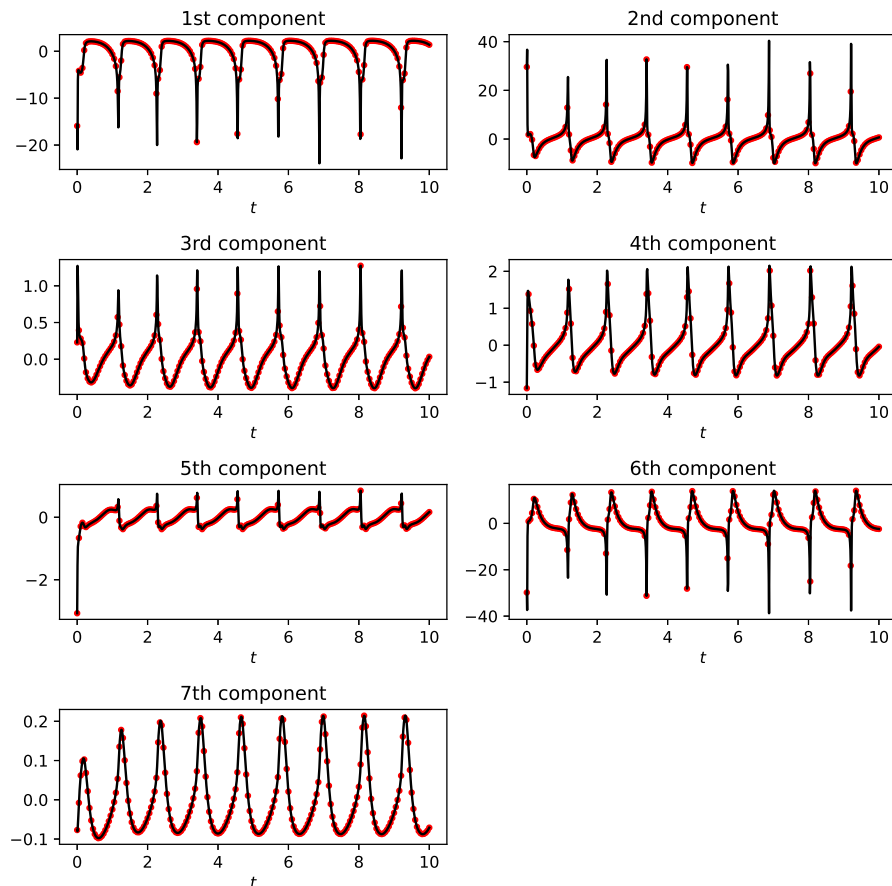


Fig. 6.10: The true governing function (solid black curves) and the approximate neural network (red circles) of Glycolytic oscillator (6.8).

- Neural Information Processing Systems*, volume 32, pages 10331–10341. Curran Associates, Inc., 2019.
- [2] Z. Allen-Zhu, Y. Li, and Z. Song. On the convergence rate of training recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 32, pages 6676–6688. Curran Associates, Inc., 2019.
  - [3] P. Amodio and L. Brugnano. The conditioning of toeplitz band matrices. *Mathematical and Computer Modelling*, 23(10):29–42, 1996.
  - [4] K. Atkinson, W. Han, and D. E. Stewart. *Numerical solution of ordinary differential equations*. Wiley, 2011.
  - [5] A. R. Barron. Neural net approximation. In *Proceedings of the 7th Yale Workshop on Adaptive and Learning Systems*, pages 69–72. Yale University Press, 1992.
  - [6] S. L. Brunton, J. L. Proctor, and J. N. Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, 2016.
  - [7] Y. Cao, Z. Fang, Y. Wu, D.-X. Zhou, and Q. Gu. Towards understanding the spectral bias of deep learning. *CoRR*, arXiv:1912.01198, 2019.
  - [8] Z. Chen, Y. Cao, D. Zou, and Q. Gu. How much over-parameterization is sufficient to learn deep relu networks? *arXiv e-prints*, arXiv:1911.12360, 2019.
  - [9] Germund Dahlquist. Convergence and stability in the numerical integration of ordinary differential equations. *Mathematica Scandinavica*, pages 33–53, 1956.
  - [10] Germund G Dahlquist. A special stability problem for linear multistep methods. *BIT Numerical Mathematics*, 3(1):27–43, 1963.
  - [11] B. C. Daniels and I. Nemenman. Efficient inference of parsimonious phenomenological models of cellular dynamics using s-systems and alternating regression. *PloS one*, 10:e0119821, 2015.
  - [12] S. S. Du, X. Zhai, B. Poczós, and A. Singh. Gradient descent provably optimizes over-parameterized neural

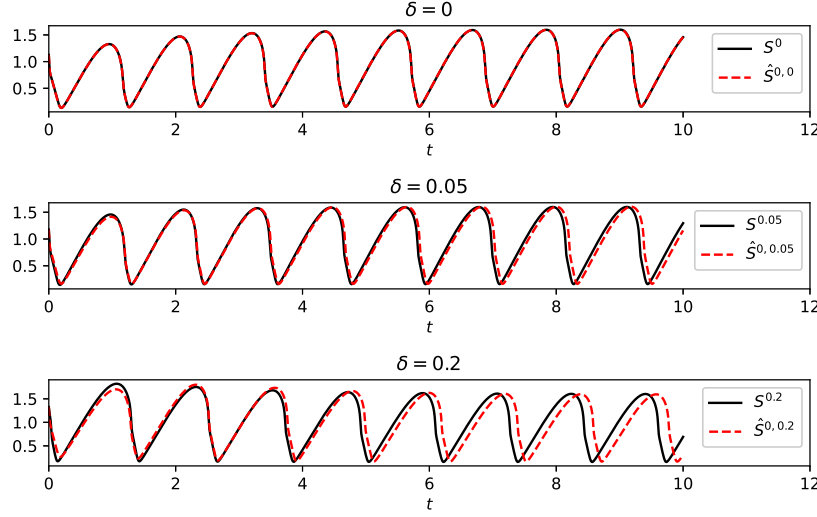


Fig. 6.11: The trajectories  $\mathbf{S}^\delta$  of exact dynamics (6.8) (solid black curves) and the trajectories  $\hat{\mathbf{S}}^{0,\delta}$  of discovered dynamics (dashed red curves) for initial values  $\mathbf{S}_0 + \delta$ ,  $\delta = 0, 0.05, 0.2$ .

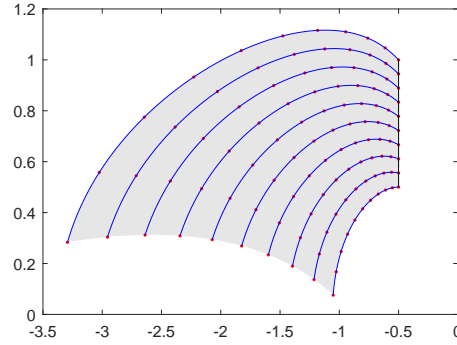


Fig. 6.12: State points (red points) collected from a sequence of trajectories (blue curves), which are computed from the model system (6.10) with initial points equidistantly distributed on the line segment  $\Gamma$  (black)

- networks. *arXiv e-prints*, arXiv:1810.02054, 2018.
- [13] W. Gautschi. *Numerical analysis*. Springer Science & Business Media, 1997.
  - [14] N. S. Gulgec, Z. Shi, N. Deshmukh, S. Pakzad, and M. Takáč. Fd-net with auxiliary time steps: fast prediction of pdes using hessian-free trust-region methods. *arXiv e-prints*, arXiv:1910.12680, 2019.
  - [15] John Harlim, Shixiao W. Jiang, Senwei Liang, and Haizhao Yang. Machine learning for prediction with missing dynamics. *Journal of Computational Physics*, 428:109922, 2021.
  - [16] Peter Henrici. *Discrete variable methods in ordinary differential equations*. Wiley, 1962.
  - [17] A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 8571–8580. Curran Associates, Inc., 2018.
  - [18] S. H. Kang, W. Liao, and Y. Liu. Ident: Identifying differential equations with numerical time evolution. *arXiv e-prints*, arXiv:1904.03538, 2019.
  - [19] R. Keller and Q. Du. Discovery of Dynamics using Linear Multistep Methods. *SIAM J. Numer. Anal.*, 59:429–455, 2021.

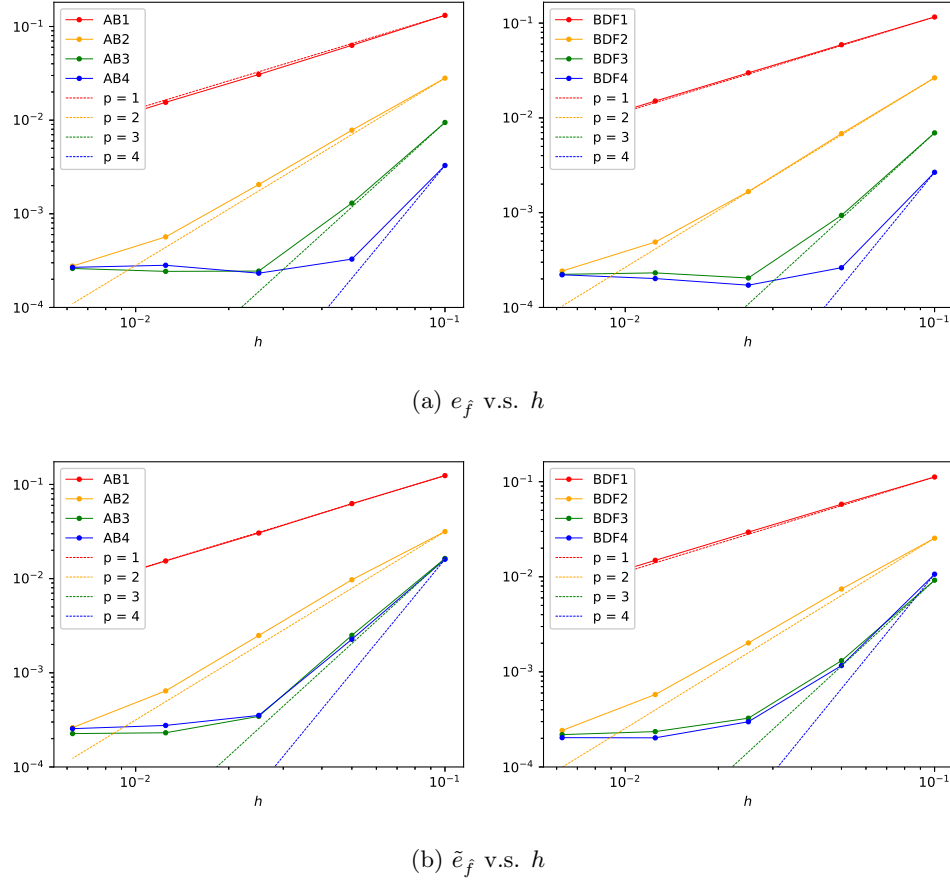


Fig. 6.13: Training error  $e_f$  and testing error  $\tilde{e}_f$  versus  $h$  via network-based A-B/BDF schemes of the model system (6.10)

- [20] Yoshua Bengio Kenji Kawaguchi, Leslie Pack Kaelbling. Generalization in deep learning. *arXiv e-prints*, arXiv:1710.05468, 2017.
- [21] D. P. Kingma and J. Ba. Adam: a method for stochastic optimization. *arXiv e-prints*, arXiv:1412.6980, 2014.
- [22] J. Kocijan, A. Girard, B. Banko, and R. Murray-Smith. Dynamic systems identification with gaussian processes. *Mathematical and Computer Modelling of Dynamical Systems*, 11:411–424, 2005.
- [23] D. Lei, Z. Sun, Y. Xiao, and W. Y. Wang. Implicit regularization of stochastic gradient descent in natural language processing: Observations and implications. *arXiv e-prints*, arXiv:1811.00659, 2018.
- [24] Z. Li, J. Han, W. E, and Q. Li. On the curse of memory in recurrent neural networks: approximation and optimization analysis. *arXiv e-prints*, arXiv:2009.07799, 2020.
- [25] Z. Long, Y. Lu, and B. Dong. "pde-net 2.0: Learning pdes from data with a numeric-symbolic hybrid deep network". *Journal of Computational Physics*, 399:108925, 2019.
- [26] F. Lu, M. Zhong, and S. Tang. Nonparametric inference of interaction laws in systems of agents from trajectory data. *Proceedings of the National Academy of Sciences*, 116(29):14424–14433, 2019.
- [27] J. Lu, Z. Shen, H. Yang, and S. Zhang. Deep network approximation for smooth functions. *arXiv e-prints*, arXiv:2001.03040, 2020.
- [28] T. Luo, Z. Ma, Z. J. Xu, and Y. Zhang. Theory of the frequency principle for general deep neural networks. *arXiv e-prints*, arXiv:1906.09235, 2019.
- [29] T. Luo and H. Yang. Two-layer neural networks for partial differential equations: Optimization and generalization theory. *arXiv e-prints*, arXiv:2006.15733, 2020.
- [30] D. Mayers and E. Süli. *An introduction to numerical analysis*. Cambridge University Press, 2003.
- [31] S. Mei, T. Misiakiewicz, and A. Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. *arXiv e-prints*, arXiv:1902.06015, 2019.

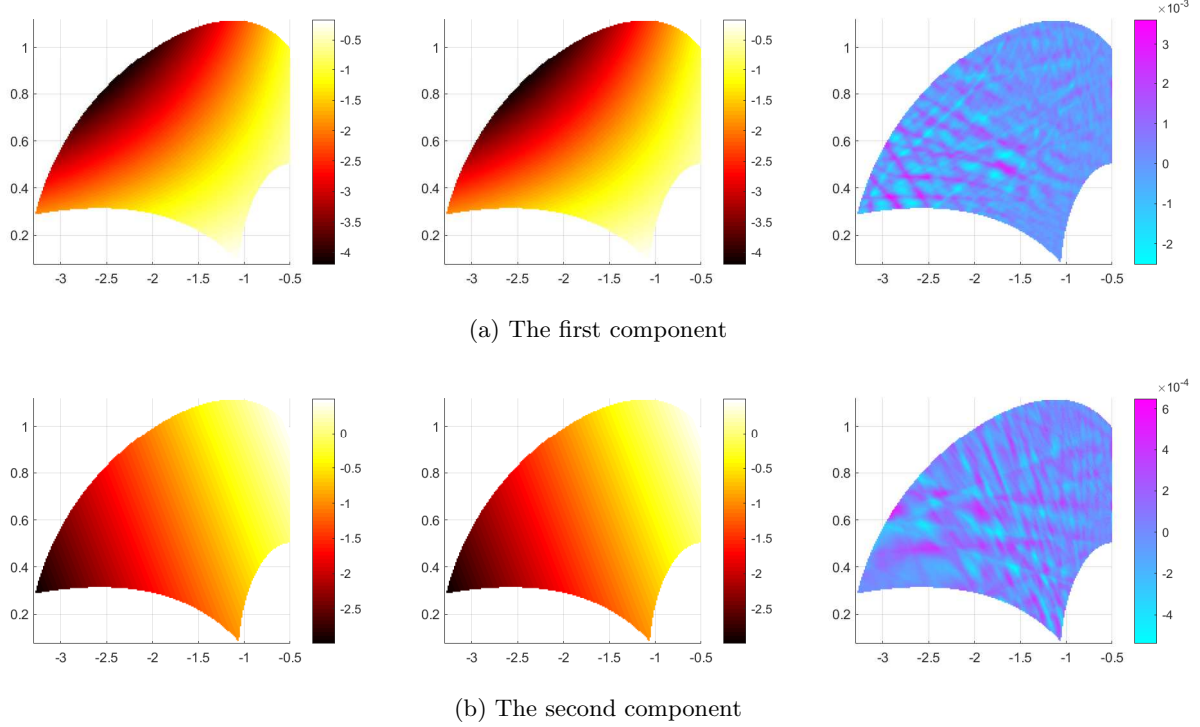


Fig. 6.14: The profiles of obtained networks (left), the true governing function (middle) and their errors (right) of the model system (6.10)

- [32] S. Mei, A. Montanari, and P.-M. Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.
- [33] B. Neyshabur, R. Tomioka, R. Salakhutdinov, and N. Srebro. Geometry of optimization and implicit regularization in deep learning. *arXiv e-prints*, arXiv:1705.03071, 2017.
- [34] S. Oymak. Stochastic gradient descent learns state equations with nonlinear activations. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2551–2579, Phoenix, USA, 25–28 Jun 2019. PMLR.
- [35] T. Qin, K. Wu, and D. Xiu. Data driven governing equations approximation using deep neural networks. *Journal of Computational Physics*, 395:620–635, 2019.
- [36] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5301–5310, Long Beach, California, USA, 09–15 Jun 2019. PMLR.
- [37] M. Raissi. Deep hidden physics models deep learning of nonlinear partial differential equations. *The Journal of Machine Learning Research*, 19(1), 2018.
- [38] M. Raissi and G. E. Karniadakis. Hidden physics models: Machine learning of nonlinear partial differential equations. *Journal of Computational Physics*, 357:125–141, 2018.
- [39] M. Raissi, P. Perdikaris, and G. E. Karniadakis. Inferring solutions of differential equations using noisy multi-fidelity data. *Journal of Computational Physics*, 335:736–746, 2017.
- [40] M. Raissi, P. Perdikaris, and G. E. Karniadakis. Machine learning of linear differential equations using gaussian processes. *Journal of Computational Physics*, 348:683–693, 2017.
- [41] M. Raissi, P. Perdikaris, and G. E. Karniadakis. Multistep neural networks for data-driven discovery of nonlinear dynamical systems. *arXiv e-prints*, arXiv:1801.01236, 2018.
- [42] A. M. Reppen and H. M. Soner. Bias-variance trade-off and overlearning in dynamic decision problems. *arXiv e-prints*, arXiv:2011.09349, 2020.
- [43] S. H. Rudy, S. L. Brunton, J. L. Proctor, and J. N. Kutz. Data-driven discovery of partial differential equations. *Science Advances*, 3(4):e1602614, 2017.
- [44] S. H. Rudy, J. N. Kutz, and S. L. Brunton. Deep learning of dynamics and signal-noise decomposition with

- time-stepping constraints. *Journal of Computational Physics*, 396:483–506, 2019.
- [45] M. Schmidt and H. Lipson. Distilling free-form natural laws from experimental data. *Science*, 324:81–85, 2009.
  - [46] Z. Shen, H. Yang, and S. Zhang. Deep network approximation characterized by number of neurons. *arXiv e-prints*, arXiv:1906.05497, 2019.
  - [47] Z. Shen, H. Yang, and S. Zhang. Deep network with approximation error being reciprocal of width to power of square root of depth. *arXiv e-prints*, arXiv:2006.12231, 2020.
  - [48] Z. Shen, H. Yang, and S. Zhang. Neural network approximation: Three hidden layers are enough. *arXiv e-prints*, arXiv:2010.14075, 2020.
  - [49] Y. Sun, L. Zhang, and H. Schaeffer. "neupde: Neural network based ordinary and partial differential equations for modeling time-dependent data". In *Proceedings of The First Mathematical and Scientific Machine Learning Conference*, volume 107 of *Proceedings of Machine Learning Research*, pages 352–372, Princeton University, Princeton, NJ, USA, 2020. PMLR.
  - [50] R. Tipireddy, P. Perdikaris, P. Stinis, and A. Tartakovsky. A comparative study of physics-informed neural network models for learning unknown dynamics and constitutive relations. *arXiv e-prints*, arXiv:1904.04058, 2019.
  - [51] S. Vaidyanathan, C. Volos, and V.-T. Pham. Hyperchaos, adaptive control and synchronization of a novel 5-d hyperchaotic system with three positive lyapunov exponents and its spice implementation. *Archives of Control Sciences*, 24:409–446, 2014.
  - [52] X. Wang and M. Wang. A hyperchaos generated from lorenz system. *Physica A: Statistical Mechanics and its Applications*, 387:3751–3758, 2008.
  - [53] K. Wu and D. Xiu. Data-driven deep learning of partial differential equations in modal space. *Journal of Computational Physics*, 408:109307, 2020.
  - [54] X. Xie, G. Zhang, and C. G. Webster. Non-intrusive inference reduced order model for fluids using deep multistep neural network. *Mathematics*, 7(8), 2019.
  - [55] Z. J. Xu, Y. Zhang, and Y. Xiao. Training behavior of deep neural network in frequency domain. In *Neural Information Processing*, pages 264–274. Springer International Publishing, 2019.
  - [56] Z. Song Z. A.-Zhu, Y. Li. A convergence theory for deep learning via over-parameterization. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 242–252, Long Beach, California, USA, 2019. PMLR.
  - [57] S. Zhang and G. Lin. Robust data-driven discovery of governing physical laws with error bars. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 474(2217):20180305, 2018.
  - [58] S. Zhang and G. Lin. Substbr to tackle high noise and outliers for data-driven discovery of differential equations. *Journal of Computational Physics*, 428:109962, 2021.

**Appendix A. Supplementary Results on Unstable LMMs.** Recall that  $\kappa_2(\mathbf{A}_h)$  denotes the 2-condition number of the matrix  $\mathbf{A}_h$  corresponding to certain LMM schemes. It has been shown in Theorem 5.4 and [19] that as  $N \rightarrow \infty$ ,  $\kappa_2(\mathbf{A}_h)$  is uniformly bounded for stable schemes. Similar arguments also show that  $\kappa_2(\mathbf{A}_h)$  increases linearly for marginally stable schemes and increases exponentially for unstable schemes. Although there has been no convergence theory for unstable schemes, it is intriguing to investigate how they perform in practice.

We first consider the discovery via linear system (3.10), in which the target function is approximated by grid functions. Note that  $\mathbf{A}_h$  is a Toeplitz-type band matrix, and hence (3.10) are linear difference equations. For unstable schemes, the characteristic polynomial has roots of modulus greater than 1, which causes small perturbations of the system to grow exponentially in the solution. Specifically, let us consider the following perturbed system of (3.10),

$$(A.1) \quad \mathbf{A}_h(\vec{f}_h + \vec{\epsilon}) = \begin{bmatrix} \mathbf{c}_h + \boldsymbol{\delta} \\ \vec{q}_h \end{bmatrix},$$

where  $\boldsymbol{\delta}$  is a small perturbation of the initial value  $\mathbf{c}_h$ , and  $\vec{\epsilon}$  is the error between the perturbed and original solutions. Then  $\vec{\epsilon} = [\epsilon_s, \epsilon_{s+1}, \dots, \epsilon_{e(N)}]^T$  is given by

$$(A.2) \quad \epsilon_n = c_1 \lambda_1^n + c_2 \lambda_2^n + \dots + c_{N_a} \lambda_{N_a}^n, \quad \forall n = s, s+1, \dots, e(N),$$

where  $\lambda_1, \dots, \lambda_{N_a}$  are the roots of the polynomial (5.28), and  $c_1, \dots, c_{N_a}$  are completely determined



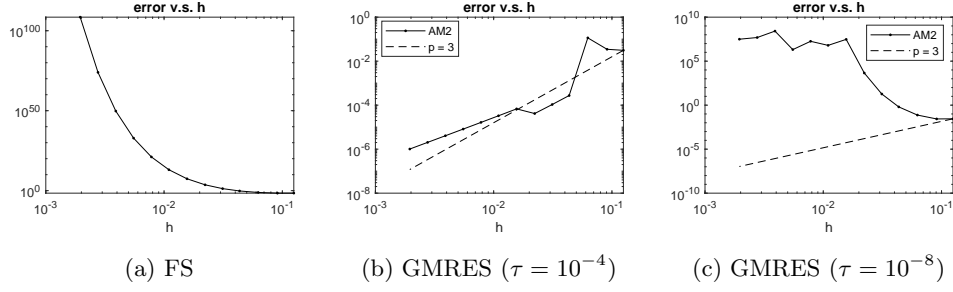


Fig. A.1: *Discovery error versus  $h$  using forward substitution (FS), GMRES ( $\tau = 10^{-4}$ ) or GMRES ( $\tau = 10^{-8}$ ) in the discovery of the model problem (6.5).*

by  $\delta$ . For unstable schemes, at least one root  $\lambda$  has modulus greater than 1, and hence the error component  $\varepsilon_n$  grows exponentially as  $n$  increases.

In practice, since  $\mathbf{A}_h$  is lower-triangular, it is natural to solve (3.10) by forward substitution directly. However, the error accumulation discussed above occurs in the process of forward substitution. To demonstrate this, we solve the linear system (3.10) concerning the unstable A-M scheme ( $M = 2$ ) to discover the dynamical system (6.5). We first use forward substitution and compute the relative discovery error

$$(A.3) \quad \left\| \vec{f}_h - \vec{f} \right\|_2 / \left\| \vec{f} \right\|_2,$$

where  $\vec{f}_h$  is the computed solution of the linear system and  $\vec{f}$  defined by (5.9) is the true governing function evaluated at grid points. It shows in Figure A.1 that the discovery error increases rapidly as  $h$  decreases, implying the failure of forward substitution.

We then repeat the test by employing iterative solvers such as the generalized minimal residual method (GMRES) with stopping residual  $\tau = 10^{-4}$ . It shows in Figure A.1 GMRES with this setting succeeds in obtaining decaying errors as  $h$  decreases, whose orders are close to the theoretical ones [19]. However, if we set a smaller stopping residual  $\tau = 10^{-8}$ , GMRES also fails like the forward substitution. Similar results are observed when using biconjugate gradient method to solve the linear system. These comparative tests imply that the difficulty brought by unstable schemes can be lessened by using iterative solvers, but these solvers are still sensitive to the implementation parameters because of the ill-conditioning of the method.

Next, we consider the discovery using linear approximation forms. Suppose the approximation set  $\mathcal{A}$  is a linear space with basis  $\{\psi_1, \dots, \psi_{d'}\}$ , then the governing function can be approximated by the form

$$(A.4) \quad \hat{f}_{\mathcal{A}} = c_1 \psi_1 + \dots + c_{d'} \psi_{d'}$$

with coefficients  $c_1, \dots, c_{d'}$  to be determined. Note that  $\mathcal{A}$  can be spaces of polynomials, finite elements, splines, etc. Under the LMM framework, we aim to compute  $c_1, \dots, c_{d'}$  such that

$$(A.5) \quad \mathbf{A}_h \mathbf{\Psi}_h \vec{c} = \begin{bmatrix} c_h \\ \vec{q}_h \end{bmatrix},$$

where  $\mathbf{\Psi}_h := [\psi_i(\mathbf{x}_n)]_{n=s, \dots, e(N)}^{i=1, \dots, d'}$  and  $\vec{c} := [c_1, \dots, c_{d'}]^T$ . Note that (A.5) is a linear system similar to (3.10) but might be square if  $d' = t(N)$ , overdetermined if  $d' < t(N)$  or underdetermined if

$d' > t(N)$ . It is natural to solve (A.5) by first solving

$$(A.6) \quad \mathbf{A}_h \vec{\mathbf{y}} = \begin{bmatrix} \mathbf{c}_h \\ \vec{\mathbf{q}}_h \end{bmatrix}$$

for  $\vec{\mathbf{y}}$ , then solve  $\Psi_h \vec{\mathbf{c}} = \vec{\mathbf{y}}$  for  $\vec{\mathbf{c}}$ . However, solving (A.6) faces the same issue as the linear system (3.10) discussed above.

Therefore, it implies that with unstable LMM schemes, both grid function approximation and linear form approximation are less robust due to the ill-conditioning. One might attempt to overcome such difficulties by developing effective preconditioners for the linear system (3.10) or (A.5), at least when there is no high demand on the numerical precision. In comparison, the network approximation shows more robustness in practice to get solutions within the ranges of optimization errors (Section 6.1.4), which is conjectured to be a consequence of All these attempts and conjectures may be further studied in future work.

