SIAM J. NUMER. ANAL.
Vol. 59, No. 1, pp. 429–455

DISCOVERY OF DYNAMICS USING LINEAR MULTISTEP METHODS*

RACHAEL T. KELLER† AND QIANG DU†

Abstract. Linear multistep methods (LMMs) are popular time discretization techniques for the numerical solution of differential equations. Traditionally they are applied to solve for the state given the dynamics (the forward problem), but here we consider their application for learning the dynamics given the state (the inverse problem). This repurposing of LMMs is largely motivated by growing interest in data-driven modeling of dynamics, but the behavior and analysis of LMMs for discovery turn out to be significantly different from the well-known, existing theory for the forward problem. Assuming a highly idealized setting of being given the exact state with a zero residual of the discrete dynamics, we establish for the first time a rigorous framework based on refined notions of consistency and stability to yield convergence using LMMs for discovery. When applying these concepts to three popular M-step LMMs, the Adams–Bashforth, Adams–Moulton, and backward differentiation formula schemes, the new theory suggests that Adams–Bashforth for M ranging from 1 and 6, Adams–Moulton for $M = 0$ and $M = 1$, and backward differentiation formula for all positive M are convergent, and, otherwise, the methods are not convergent in general. In addition, we provide numerical experiments to both motivate and substantiate our theoretical analysis.

Key words. discovery of dynamics, data-driven modeling, linear multistep methods, stability and convergence, root condition, learning dynamics, artificial intelligence

AMS subject classifications. 65L06, 65L09, 65L20, 65P99, 68T99

DOI. 10.1137/19M130981X

1. Introduction.

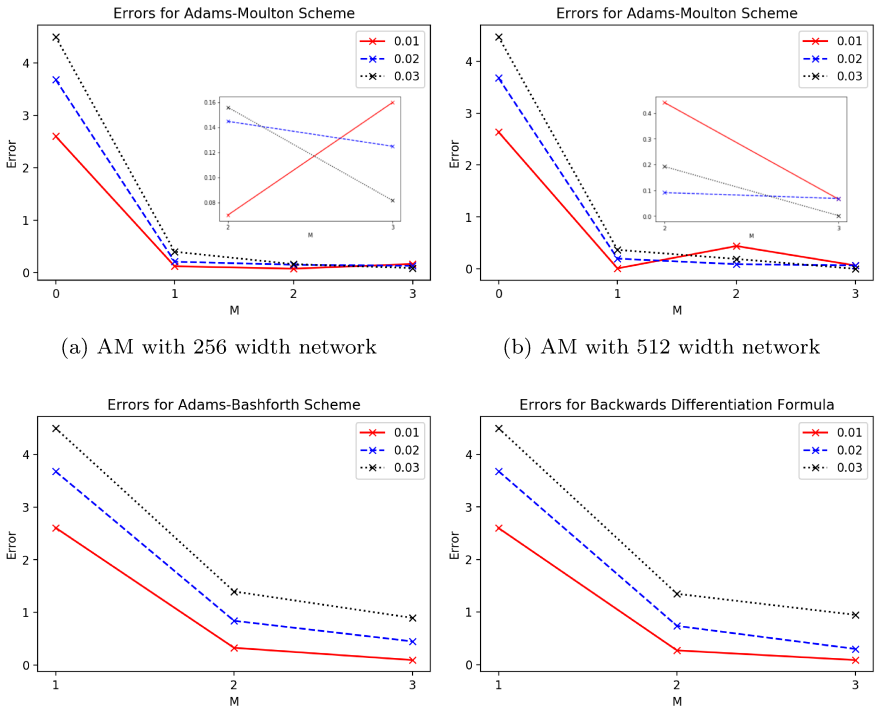
In this work, we focus on developing a new numerical analysis framework for the *discovery* of dynamical systems with given states, where finitely many discrete measurements are used to approximately recover the unknown dynamical system—a *data-driven* discovery of dynamics [5, 46]. Data-driven discovery of dynamical systems is experiencing a renaissance as costs of sensors, data storage, and computational resources has decreased [44]. Meanwhile, advancements in the fields of machine learning and data science [17, 23, 28, 29, 47] have brought in renewed vigor and enabled an expansive view of this field. At the same time, the growth of data-driven discovery of dynamical systems has also led to a new solution method and model reduction approach to study multiscale and high dimensional complex problems. For more discussions, we refer to works such as [3, 6, 18, 20, 24, 26, 27, 30, 31, 32, 37, 38, 39, 40, 41, 42, 43, 45, 50, 52, 53, 55, 56, 57, 58].

1.1. Motivation: Data-driven discovery of dynamical systems via linear multistep methods.

In this work, we consider using linear multistep methods (LMMs) to discover unspecified dynamics given the state at equidistant time steps and contribute to the fundamental theory of using LMMs for data-driven discovery. Historically, LMMs have been developed as popular schemes for numerically integrating known dynamic systems [16], with well-established mathematical theory in the last

*Received by the editors December 30, 2019; accepted for publication (in revised form) October 20, 2020; published electronically February 18, 2021.
https://doi.org/10.1137/19M130981X
Funding: The work of the first author was supported by the National Science Foundation Graduate Research Fellowship grant DGE-1644869. The work of the second author was supported by the National Science Foundation grants DMS-1719699, DMS-2012562, CCF-1704833, and ARO MURI grant W911NF-15-1-0562.

†Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY 10027 USA (rachael.keller@columbia.edu, qd2125@columbia.edu).



(a) AM with 256 width network

(b) AM with 512 width network

(c) AB with 256 width network

(d) BDF with 256 width network

FIG. 1. *Absolute ℓ_2-errors for the first coordinate of the 2D Damped Cubic System* (6.1) *on $t \in [0, 5]$ with varying time mesh size $h = 0.01, 0.02, 0.03$, using a single hidden layer neural network with* tanh *activation function, as used in* [41]*, after a fixed number of training iterations for each M.*

century [2, 12, 15, 22, 33]. Recent works combine the classical numerical technique of linear multistep methods with neural networks for dynamics discovery [41, 52, 57].

Coined "LMNet," LMMs combined with neural networks are used for the discovery of dynamics in [41, 52, 57]. Figure 1 shows the absolute errors associated with learning $\boldsymbol{f}$ for a nonlinearly damped, two-dimensional (2D) cubic oscillator (6.1) using neural networks with three representative schemes of LMMs—Adams–Moulton (AM), Adams–Bashforth (AB), and backward differentiation formula (BDF). These results are generated using the code repository built for [41]; reported are the errors of the dynamics rather than the integrated dynamics, which are shown in [41]. For solving differential equations with smooth solutions, increasing M corresponds to higher accuracy if the scheme is also stable. The AM scheme is an example of such a method, hence, the perplexing behavior in the errors of AM as observed in [41, 57] (see Tables 1 and 2 of [41] and Table 1 of [57]). As M increases and h decreases, the errors do not decrease. Further, as we expand the width, thereby increasing the expressibility of the network, the scheme still does not exhibit stable behavior. On the other hand, as shown in Figure 1, the AB and BDF methods with a fixed network size of 256 show a trend of convergence as M and the mesh size h decrease, while the AM methods show erratic behavior for the same width, persistent even with more expressibility of the network by widening the hidden layer (Figure 1(b)). Since AM is a stable method as a time integrator, these findings warrant further investigation. Indeed, it has also been

observed by others that increased resolution does not necessarily imply better neural network representation and prediction without a mathematically sound formulation of the learning problem [3]. While there are many contributing factors, such as the neural network structure and size as well as the training process, it is the goal of this paper to investigate these findings and provide a theoretical explanation of the phenomena.

To begin, we pose the problem of discovery of dynamics. In contrast to numerically integrating dynamics to learn the state, as many classical numerical methods do, this study focuses on learning the dynamics given the state. Dynamics discovery may be viewed as an inverse problem to the forward problem of classical numerical integration. Well-studied for the forward problem, LMMs in this inverse setting raise questions of classical notions of consistency, stability, and convergence. We seek in this work to investigate if the classical theory for LMMs as time integrators to solve the forward problem has an analogue or counterpart in solving the inverse problem of learning dynamics. To initiate studies in this direction, we introduce a systematic framework for the numerical analysis of the discovery of dynamics using LMMs. Our new framework is rooted in the classical theory for LMMs as numerical integrators of differential equations, but it adopts new stability and convergence criteria due to the inverse nature of using discrete time integrators for dynamics discovery. Consequently, it draws different conclusions regarding convergence in stark contrast to the conventional wisdom. The stability properties of particular schemes depart from the traditional numerical differential equation viewpoint, and some methods that are stable for the forward problem do not retain the property for the inverse problem dynamics discovery. Our theory is able to explain the unusual phenomena as reported in Figure 1 and lays a rigorous foundation for further elucidating the effect of neural networks on dynamics discovery via LMMs through follow-up studies. Therefore, this helps the scientific community broadly in our goal of making machine learning more transparent, explainable, stable, and trustworthy.

1.2. Summary of results. We present a framework in section 3 consisting of nuanced notions of consistency and stability to handle unique challenges presented by using LMMs for discovery. These concepts are then combined to prove convergence. A set of algebraic criteria is developed to check for the consistency and stability, and thus convergence, of LMMs for dynamics discovery. With this foundation, in Theorems 4.1 and 4.2, we outline consistency and stability properties of the AB, AM, and BDF schemes, and consequentially, in Corollary 4.3, their convergence guarantees.

1.3. Outline. This paper is organized as follows. In section 2 we briefly review LMMs and their theory for solving ordinary differential equations, including the standard notions in numerical analysis of truncation error, consistency, stability, and convergence, along with an algebraic root condition for stability. In section 3 we frame the problem of discovery using LMMs and develop nuanced versions of consistency and stability for discovery. In particular, in section 3.3, we discuss how truncation error for discovery is inherited from the forward problem and introduce a stronger notion of consistency; in section 3.4 we refine the traditional definition of stability and the algebraic root condition, and we show equivalent theorems connecting the root conditions and the refined notions of stability. In section 4, the discovery framework of section 3 is applied to characterize convergence properties of the AB, AM, and BDF schemes. Some discussions on the long-time dynamics discovery are made in section 5. Then, in section 6, we show results of numerical experiments. Finally, in section 7, we summarize the results and discuss future directions.

2. LMMs: Quick review. In this section, we introduce notation used throughout this work and briefly review the theory of LMMs as time integrators. While LMMs are well-documented in standard textbooks for solving ordinary differential equations (see [15, 33, 2, 22]), we include the salient points to facilitate direct comparison with the new theory for the discovery of unknown dynamics developed in the next section.

2.1. LMMs: Notation and concepts. Consider the ordinary differential equation (ODE)

$$\frac{d}{dt}\mathbf{x}(t) = \boldsymbol{f}(\mathbf{x}(t)),\ a \le t \le b,\ \mathbf{x}(t_0) = \mathbf{x}_0, \tag{2.1}$$

where $\mathbf{x} \in C^\infty(0,\infty)^d$ and f is assumed to be a Lipschitz continuous, smooth, and bounded function. Discretizing the model problem (2.1), we assume a grid on the interval $[\,a,b\,]$ defined to be a set of points: $a = t_0 < t_1 < \cdots < t_N = b$ with equidistant mesh $t_{n+1} - t_n = h = (b-a)/(N+1)$, $n \in \{0,1,\ldots,N\}$. Let $[\,a,b\,]_h$ denote this ordered set. We denote the set of grid functions $\Gamma_h[\,a,b\,] = \left\{\mathbf{z} \,|\, \mathbf{z} \in \mathbb{R}^{(N+1)\times d},\ \mathbf{z}_n = \mathbf{z}(t_n) \in \mathbb{R}^d, t_n \in [\,a,b\,]_h\right\}$ [15].

An M-step LMM approximates the nth value $\mathbf{x}_n = \mathbf{x}(t_n)$ in terms of the previous M ($M \ge 1$) time steps $\mathbf{x}_{n-1}, \mathbf{x}_{n-2}, \ldots, \mathbf{x}_{n-M}$ [15, 33, 2, 22]. An M-step linear multistep method is given by

$$\sum_{m=0}^{M} \alpha_m \mathbf{x}_{n-m} = h \sum_{m=0}^{M} \beta_m \boldsymbol{f}(\mathbf{x}_{n-m}),\ n = M, M+1, \ldots, N, \tag{2.2}$$

where $\mathbf{x} \in \Gamma_h[\,a,b\,]$, the coefficients $\alpha_m, \beta_m \in \mathbb{R}$ for $m = 0,1,\ldots,M$, and $\alpha_0 \neq 0$. The function $\boldsymbol{f}$ is assumed to be given and Lipschitz, and the LMM scheme (2.2) defines an iterative procedure stepping forward in the independent variable $t \in [\,a,b\,]$ to solve for $\mathbf{x}(t)$ at the gridpoints. Associated with an M-step LMM are its first and second characteristic polynomials, given, respectively, by

$$\rho(z) = \sum_{m=0}^{M} \alpha_{M-m} z^m, \quad \text{and} \quad \sigma(z) = \sum_{m=0}^{M} \beta_{M-m} z^m, \tag{2.3}$$

where it is assumed that $\alpha_0 \neq 0$ [33].

For the numerical integration of differential equations, the method (2.2) is called explicit if $\beta_0 = 0$ and implicit otherwise [15, 33, 2]. Implicit methods require a nonlinear solver to the generated system of equations, whereas explicit methods do not. The existence and uniqueness of solutions in the case of implicit schemes is shown in [15, 22]. For both implicit and explicit methods, a kickstarting method for initial M values must be chosen, and as such a critical component of analyzing any multistep method scheme is to understand how much errors in initial values pollute the subsequent calculations [15]. This aspect of numerical methods is called numerical stability [2].

Finally, for any index set $\mathcal{S}$ with cardinality $\bar{S}$, we let $\|\mathbf{z}\|_1 = \sum_{i\in\mathcal{S}} |\mathbf{z}_i|$ and $\|\mathbf{z}\|_\infty = \max_{i\in\mathcal{S}} |\mathbf{z}_i|$ denote the standard discrete norms for any vector $\mathbf{z}$ naturally embedded in $\mathbb{R}^{\bar{S}\times d}$ where $|\mathbf{z}_i|$ can be any vector norm of $\mathbf{z}_i \in \mathbb{R}^d$. The same notation is used also for discrete grid functions given either in $\Gamma_h[a,b]$ or its subsets.

Remark 2.1. To fix ideas, we use the hat notation $\hat{}$ to mark grid functions generated by the discretization (2.2). In the forward problem, the state $\mathbf{x}(t)$ is iteratively produced by LMMs, and hence we study $\hat{\mathbf{x}}$, whereas for dynamics discovery, we study $\hat{\boldsymbol{f}}$ (see section 3).

2.2. The Adams family and BDF. AB, AM, and the BDF are three popular multistep method schemes that arise from a Lagrange interpolating polynomial of the state or dynamics at time t_n using data from previous time steps. Without loss of generality, we consider the scalar model problem in this section; for higher dimensions, the theory need only be applied in each dimension. Let $\Lambda_0 = \{-M, -M+1, \ldots, -1, 0\}$ and $\Lambda_1 = \{-M, -M+1, \ldots, -1\}$. The Lagrange interpolating polynomial of a function $u : \mathbb{R} \to \mathbb{R}$ over the set $\{t_{n+i}, i \in \tilde{\Lambda}\}$ is the polynomial of degree M for $\tilde{\Lambda} = \Lambda_0$ and degree $M-1$ for $\tilde{\Lambda} = \Lambda_1$ obtained from the linear combination of basis functions

$$\ell_{k,n}(t; \tilde{\Lambda}) = \prod_{i \in \tilde{\Lambda} \setminus \{k\}} \frac{t - t_{n+i}}{t_{n+k} - t_{n+i}}, \ k \in \tilde{\Lambda}, \tag{2.4}$$

with $u(t_{n+k})$ for each $k \in \tilde{\Lambda}$ as the coefficient of the linear combination. The M-step AM (or AM-M) and AB (or AB-M) are M-step LMMs that arise from interpolating the dynamics $f(x(t_n))$, with Lagrange interpolating polynomials corresponding to $\tilde{\Lambda} = \Lambda_0$ and $\tilde{\Lambda} = \Lambda_1$, respectively, and then applying the fundamental theorem of calculus on the model problem (2.1). Letting $f(t_n)$ denote $f(x(t_n))$ for brevity, we have

$$x(t_n) \approx x(t_{n-1}) + \int_{t_{n-1}}^{t_n} \sum_{k \in \tilde{\Lambda}} f(t_{n+k}) \ell_{k,n}(t; \tilde{\Lambda}) dt \,. \tag{2.5}$$

BDF-M, on the other hand, is an M-step LMM for $M \geq 1$ derived from interpolating the state $\mathbf{x} \in \Gamma_h[a, b]$ in (2.1) directly on the lattice Λ_0, so that

$$\sum_{k \in \Lambda_0} x(t_{n+k}) \frac{d\ell_{k,n}}{dt}(t_n; \Lambda_0) \approx \frac{d}{dt} x(t_n) = f(t_n).$$

By the change of variables $u = (t - t_{n-1})/h$, we have a scaled Lagrange interpolating polynomial, denoted ℓ_k^h, given by

$$\ell_k^h(u; \tilde{\Lambda}) = \prod_{i \in \tilde{\Lambda} \setminus \{k\}} \frac{u - 1 - i}{k - i}, \ k \in \tilde{\Lambda}. \tag{2.6}$$

With (2.6), the integrand of (2.5) may be written independently of the time step, so that

$$x(t_n) \approx x(t_{n-1}) + \int_0^1 \sum_{k \in \tilde{\Lambda}} f(t_{n+k}) \ell_k^h(u; \tilde{\Lambda}) du. \tag{2.7}$$

The simplified coefficients for the BDF method with uniform mesh can be obtained similarly.

2.3. Truncation error and consistency. In this section, we introduce the residual and notions related to analytical error for LMMs. The residual operator is given by [15]:

$$(R_h\hat{\mathbf{x}})_n := \frac{1}{h}\sum_{m=0}^{M}\alpha_m\hat{\mathbf{x}}_{n-m} - \sum_{m=0}^{M}\beta_m\boldsymbol{f}(\hat{\mathbf{x}}_{n-m}), \quad n = M, M+1, \ldots, N, \tag{2.8}$$

defined for $\hat{\mathbf{x}} \in \Gamma_h[a,b]$. How accurately the discretization (2.2) approximates the solution of (2.1) is measured by the truncation error, defined below.

DEFINITION 2.2 (local truncation error [33, 2, 22, 15]). *Let $\boldsymbol{x} \in \Gamma_h[a,b]$ be the exact solution of the dynamic system* (2.1) *defined at the grid coordinates. The local truncation error $\boldsymbol{\tau}_h = ((\boldsymbol{\tau}_h)_M, (\boldsymbol{\tau}_h)_{M+1}, \ldots, (\boldsymbol{\tau}_h)_N) \in \mathbb{R}^{(N-M+1)\times d}$ is given by*

$$(\boldsymbol{\tau}_h)_n = (R_h\boldsymbol{x})_n \quad \text{for } n = M, M+1, \ldots, N. \tag{2.9}$$

For smooth functions $\boldsymbol{f}$ and $\boldsymbol{x}$, we have

$$(\boldsymbol{\tau}_h)_n = \sum_{m=0}^{\infty} C_m h^{m-1}\nabla_t^m\boldsymbol{x}(t_n) \quad \text{for } n = M, M+1, \ldots, N,$$

where

$$C_0 = \sum_{k=0}^{M}\alpha_k,\ C_m = (-1)^m\left[\frac{1}{m!}\sum_{k=1}^{M}k^m\alpha_k + \frac{1}{(m-1)!}\sum_{k=0}^{M}k^{m-1}\beta_k\right],\ m = 1, 2, \ldots. \tag{2.10}$$

Now, we proceed to define order of error and the notion of consistency.

DEFINITION 2.3 (order of error [15]). *A linear multistep method has an error order of p if $\|\boldsymbol{\tau}_h\|_\infty = \mathcal{O}(h^p)$ as $h \to 0$ and admits a* principal error function *$\boldsymbol{e}(t) \in C[a,b]$ provided*

$$\boldsymbol{e}(t) \neq \mathbf{0} \text{ and } (\boldsymbol{\tau}_h)_n = \boldsymbol{e}(t_n)h^p + \mathcal{O}(h^{p+1}) \text{ as } h \to 0,$$

or simply, $\|\boldsymbol{\tau}_h - h^p\boldsymbol{e}\|_\infty = \mathcal{O}(h^{p+1})$.

DEFINITION 2.4 (consistency [15]). *A linear multistep method is consistent with the differential equation provided $\|\boldsymbol{\tau}_h\|_\infty \to 0$ as $h \to 0$.*

The Adams family and BDF are consistent in the sense of Definition 2.4. Moreover, the local truncation error associated with the M-step AB and BDF schemes are $\mathcal{O}(h^M)$, whereas for the M-step AM, the local truncation error is $\mathcal{O}(h^{M+1})$ [33, 2].

It is well-known that consistency can be formulated algebraically in terms of the characteristic polynomials [11]. In particular, the consistency condition, i.e., $C_0 = C_1 = 0$ in (2.10), is equivalent to $\rho(1) = 0$ and $\rho'(1) = \sigma(1)$. Moreover, the truncation error is order k if

$$\rho(e^z) - z\sigma(e^z) = O(z^{k+1}), \quad \text{as} \quad z \to 0. \tag{2.11}$$

2.4. Stability and the root condition. In this section, we review definitions of stability and the root condition for LMMs. Stability is defined as follows.

DEFINITION 2.5 (stability [15]). *A linear M-step method for the ordinary differential equation $\dot{\boldsymbol{x}} = \boldsymbol{f}(t, \boldsymbol{x}(t))$ is called stable on $[a, b]$ provided there exists a constant K not depending on h such that, for any two grid functions $\boldsymbol{u}, \boldsymbol{v} \in \Gamma_h[a, b]$, we have for all h sufficiently small*

$$\|\boldsymbol{u} - \boldsymbol{v}\|_\infty \leq K \left(\max_{0 \leq i \leq M-1} \|\boldsymbol{u}_i - \boldsymbol{v}_i\| + \|R_h \boldsymbol{u} - R_h \boldsymbol{v}\|_\infty \right).$$

The characteristic polynomials defined in (2.3) may be used to determine the stability of a linear multistep method via the root condition.

DEFINITION 2.6 (algebraic root condition [33, 15]). *A polynomial satisfies the root condition provided the roots of the polynomial do not exceed magnitude* 1*, and those of magnitude* 1 *are simple.*

The following theorem states the equivalence between the stability and the root condition.

THEOREM 2.7 (stability and the root condition [33, 15]). *A linear multistep method is stable if and only if its first characteristic polynomial $\rho(z)$ satisfies the algebraic root condition given by Definition* 2.6*.*

Note that all AB and AM schemes satisfy the root condition and are stable by Definition 2.5, whereas BDF-M satisfies the root condition and is stable only for $1 \leq M \leq 6$ [22].

2.5. Convergence. Finally, we introduce the definition of convergence for LMMs and the celebrated equivalence theorem for determining it.

DEFINITION 2.8 (convergence [15]). *Consider the initial value problem* (2.1) *and a fixed linear multistep method defined by* (2.2)*. Let $\hat{\boldsymbol{x}} \in \Gamma_h[a, b]$ be the grid function obtained by applying* (2.2) *on a uniform, real-valued grid of $[a, b]$ with mesh size h, and let $\boldsymbol{x} \in \Gamma_h[a, b]$ be the exact solution of* (2.1) *at the grid points. The linear multistep method is said to converge on $[a, b]$ if*

$$\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_\infty \to 0 \quad \textit{as} \quad h \to 0 \quad \textit{whenever} \quad \max_{0 \leq k \leq M-1} \|\hat{\boldsymbol{x}}_k - \boldsymbol{x}(t_k)\|_\infty \to 0.$$

With Definition 2.8, one can obtain the Dahlquist equivalence theorem, Theorem 2.9 [33].

THEOREM 2.9 (equivalence theorem [15]). *The multistep method* (2.2) *converges in the sense of Definition* 2.8 *for all Lipschitz $\boldsymbol{f}$ if and only if it is consistent and stable.*

From the equivalence theorem, it can be shown that the order of the error $\|\mathbf{x} - \hat{\mathbf{x}}\|_\infty$ is the same order as the truncation error (Definition 2.3) and thus the order of approximation, provided the initial error $\max_{0 \leq k \leq M-1} |\hat{\mathbf{x}}_k - \mathbf{x}(t_k)|$, is also of the same order.

In this work, we develop an analogous theory for multistep methods modifying these theorems to deal with the discovery of dynamics rather than solving the differential equation. In particular, we show how the second characteristic polynomial is determinant of stability for discovery and whether the Adams family and BDF are stable or not.

3. Discovery of dynamics. In this study, we consider a *data-driven* technique to solve for the dynamics $\boldsymbol{f}$ given information on the state $\mathbf{x}$ at equidistant time steps [41]. First, we introduce the problem and then discuss notions of consistency, stability, and convergence. We now proceed to define the problem of LMMs for discovery.

3.1. Problem definition. Following earlier discussions, we are concerned with the initial value problem (2.1). In this section and the next, multivariate functions representing the continuum models are denoted by scalar notations, i.e., $f = f(x)$ and $x = x(t)$, so that boldface symbols can be reserved for vectors corresponding to discrete forms of dynamics, which should be clear in context without ambiguity. The task of learning is to produce a function to approximately represent the dynamics, $f = f(x)$, based on a set of observed states, that conforms with the discrete dynamics described by a linear multistep method. In practice, one often encounters situations with only partial (incomplete) data or data containing observation errors and uncertainties; these complications are typical for inverse problems. When combined with deep networks, the approximation is produced by a network in a learned parametrized form, which introduces further approximations as well as implicit regularizations.

As the first step to develop a rigorous numerical analysis framework, we consider a very idealized setting in this work by assuming that (A1) a complete set of exact values of the state, $\{\mathbf{x}_n = x(t_n)\}_{n=0}^N$, given at equally distributed, ordered grid points $\{t_n\}_{n=0}^N$; (A2) the neural networks (or the underlying function classes used to represent the dynamics) have sufficient approximation capability to produce zero residual for the discrete dynamical system; and (A3) approximated values of the exact dynamics for some observed initial states are available.

Although the assumptions make the situation very idealized, the study is a very constructive step toward the understanding of the mathematical and computational issues related to the data-driven modeling using neural networks and discretized forms of the unknown dynamics, which are the focuses of our ongoing work. The findings made here shed light on future studies of similar issues under more realistic conditions, as discussed in section 3.2 and further in section 7. Under the assumptions (A1), (A2), and (A3) stated above, the procedure of learning dynamics can be described as follows. Given $\mathbf{x}_n = x(t_n)$ for $0 \le n \le N$ and $\hat{\boldsymbol{f}}_i$ as suitable approximations of $f(\mathbf{x}_i)$ for i in a suitable subset of $\{0 \le i \le M-1\}$, we have zero residuals for the discrete dynamics based on the LMM discretization for t_n with $n = M, \dots, N$, i.e.,

$$\sum_{m=0}^{M} \beta_m \hat{\boldsymbol{f}}_{n-m} = \frac{1}{h} \sum_{m=0}^{M} \alpha_m \mathbf{x}_{n-m}, \quad n = M, M+1, \dots, N.$$

Indeed, the above system for $\hat{\boldsymbol{f}}$ is simply (2.2) rewritten for learning the dynamics rather than the state. To help with later discussions, we let $N_M = N - M + 1$ denote the number of linear equations in the system. Given that the values of $\{\beta_m\}_{m=0}^M$ affect the structure of the resulting system, we let m_0 and M_0 be the smallest and the largest index, respectively, among those m's satisfying $\beta_m \neq 0$, i.e.,

$$\beta_m = 0 \ \text{ for any } m \text{ with either } m < m_0 \text{ or } m > M_0, \ \text{ while } \beta_{m_0} \neq 0 \text{ and } \beta_{M_0} \neq 0.$$

We collect the ordered coefficients of the LMM scheme in the vectors $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \dots, \alpha_M)$ and $\boldsymbol{\beta} = (\beta_{m_0}, \beta_{m_0+1}, \dots, \beta_{M_0})$. The system for $\hat{\boldsymbol{f}}$ in this reduced notation is then

$$\sum_{m=m_0}^{M_0} \beta_m \hat{\boldsymbol{f}}_{n-m} = \frac{1}{h} \sum_{m=0}^{M} \alpha_m \mathbf{x}_{n-m}, \quad n = M, M+1, \dots, N. \tag{3.1}$$

For brevity, we introduce the index sets $\mathcal{I} = \{n \in \mathbb{N} \,|\, M - m_0 \le n \le N - m_0\}$ for the set of indices of the grid associated with the values of unknown dynamics and $\mathcal{I}_M := \{n \in \mathbb{N} \,|\, M - M_0 \le n < M - m_0\}$ for the set of indices for supplied initial dynamics. The linear system (3.1) may be written in compact matrix-vector form:

$$B\hat{\boldsymbol{f}} = h^{-1}A\mathbf{x} - \hat{\mathbf{g}}, \tag{3.2}$$

where A is the $N_M \times (N+1)$ matrix of coefficients for $\boldsymbol{\alpha}$ corresponding to $\mathbf{x}_{n-m}$ in (3.1); the matrix B is an $N_M \times N_M$ banded lower-triangular matrix with its diagonal entries given by β_{m_0} and the kth subdiagonal entries given by β_{m_0+k} for $k = 1, 2, \ldots, M_0 - m_0$; $\hat{\boldsymbol{f}} \in \mathbb{R}^{N_M \times d}$ is the ordered vector of unknowns $\{\hat{\boldsymbol{f}}_n\}_{n\in\mathcal{I}}$; and $\hat{\mathbf{g}} = (\hat{\mathbf{g}}_M, \hat{\mathbf{g}}_{M+1}, \ldots, \hat{\mathbf{g}}_N) \in \mathbb{R}^{N_M \times d}$ is defined as

$$\hat{\mathbf{g}}_n = \begin{cases} \displaystyle\sum_{\substack{m \ge n - M_0 \\ m \in \mathcal{I}_M}} \beta_{n-m}\hat{\boldsymbol{f}}_m & \text{if } \ n \in M_0 + \mathcal{I}_M, \\ 0 & \text{if } \ n \in \mathcal{I} \setminus \{M_0 + \mathcal{I}_M\}, \end{cases}$$

which can be generated from the assumed, suitably approximated starting values $\{\hat{\boldsymbol{f}}_n\}_{n\in\mathcal{I}_M}$. We note that since $\beta_{m_0} \neq 0$, B^{-1} always exists so that (3.2) is solvable whenever the right-hand terms are prescribed.

3.2. Connection to machine learning-based data-driven discovery and LMNet. To see how the theory developed in this work is connected to the increasingly popular machine learning based data-driven discovery of dynamics, we briefly recall the relevant learning problems here. For more extensive works on machine learning, we refer to [4, 17, 34, 35, 47].

In a generic supervised machine learning setting of learning an unknown function $\boldsymbol{f}$, one often assumes knowledge of $\tilde{N}$ samples of input-output data, $\mathcal{D} = \{(\mathbf{x}_n, \boldsymbol{f}(\mathbf{x}_n))\}_{n=1}^{\tilde{N}}$. This sample dataset is often divided into sets of training and test sets, and one attempts to find a neural network (NN) representation of $\boldsymbol{f}$, say, $\boldsymbol{f}_{NN}$, through an empirical loss minimization over the training set. We let $\tilde{\mathbf{x}}$ and $\tilde{\boldsymbol{f}}$ denote an ordered subset of $\tilde{K} \le \tilde{N}$ data, so that $(\tilde{\mathbf{x}}_k, \tilde{\boldsymbol{f}}_k) = \big(\mathbf{x}_{n_k}, \boldsymbol{f}(\mathbf{x}_{n_k})\big) \in \mathcal{D}$. The loss is a suitably defined function $\ell(\tilde{\mathbf{x}}, \tilde{\boldsymbol{f}}, \boldsymbol{f}_{NN})$ measuring a distance between $\boldsymbol{f}(\mathbf{x}_{n_k})$ and $\boldsymbol{f}_{NN}(\mathbf{x}_{n_k})$ for each $k = 1, 2, \ldots, \tilde{K}$. When evaluated over only training data, this loss leads to the training error. The desired goal is to learn $\boldsymbol{f}_{NN}$ that not only minimizes the loss in the training set (i.e., the training error), but also achieves a small loss in the remaining test samples (i.e., the generalization error).

In the setting of dynamics discovery, it is important to note that the dynamics, or output, data is not given directly. Instead, only the state, or the input, is provided, and information on the true dynamics $\boldsymbol{f}$ is inferred by constraining the data to conform with some dynamical system. For the LMM discretization of the dynamics given by (3.1), conformity is achieved by minimizing the error associated with the LMM system, which we call the LMM residual. A total loss function of the optimization problem may be effectively viewed as

$$\mathcal{T}(\tilde{\mathbf{x}}, \tilde{\boldsymbol{f}}, \boldsymbol{f}_{NN}) = \tilde{\ell}(\tilde{\mathbf{x}}, \tilde{\boldsymbol{f}}) + \ell(\tilde{\mathbf{x}}, \tilde{\boldsymbol{f}}, \boldsymbol{f}_{NN}),$$

where the loss $\tilde{\ell}$ is an increasing function of the LMM residual and vanishes at the origin, e.g., $\tilde{\ell}(\tilde{\mathbf{x}}, \tilde{\boldsymbol{f}}) = \|B\tilde{\boldsymbol{f}} - h^{-1}A\tilde{\mathbf{x}} + \hat{\mathbf{g}}\|_2^2$. A network approximation with sufficient accuracy would attempt to conform with the discretized LMM dynamics by minimizing the LMM residual to find the unknown data $\tilde{\boldsymbol{f}}$. Alternatively, as done in LMNet,

the neural network approximation may be supplied to the LMM residual $\tilde{\ell}$, where the initial dynamics in $\hat{\mathbf{g}}$ are also learned. If the approximation can be as accurate as desired, we would be led to the idealized setting that as the network is trained more, given sufficient width, the neural network would converge to $\hat{\boldsymbol{f}}$, where $\hat{\boldsymbol{f}}$ satisfies (3.2).

Naturally, due to other practical considerations as well as the finite approximation power of the neural networks, more general loss functions, regularization techniques, and network architectures may also be taken into account; see section 7 for further discussions. Our main focus here is to illustrate the impact of using different LMMs on the learning process by developing a rigorous mathematical theory of consistency, stability, and convergence for the dynamics discovery, beginning with the highly idealized setting of exact state data.

3.3. Truncation error and consistency. LMMs for discovery inherit the truncation error of solving ordinary differential equations with LMMs. Indeed, truncation error is specific to the discretization of the continuous problem; therefore, the truncation error $\boldsymbol{\tau}_h$ of a scheme for dynamics discovery remains the same as that for solving an ordinary differential equation for the state defined by (2.9). However, in addition to inheriting the same concept of consistency from section 2, Definition 2.4, we also introduce some strengthened notions of consistency for dynamics discovery. We complement these concepts later on with refined notions of stability for a more nuanced discussion of convergence for discovery using LMMs. Consistency and its strengthened forms are defined below.

DEFINITION 3.1 (consistency for dynamics discovery). *An LMM is consistent with the differential equation for dynamics discovery provided* $\|\boldsymbol{\tau}_h\|_\infty \to 0$ *as* $h \to 0$*, and it is strongly consistent if* $\|\boldsymbol{\tau}_h\|_1 \to 0$ *as* $h \to 0$*. Furthermore, a method is consistent of degree* k*, for* $k \geq 1$*, provided* $N^{k-1}\|\boldsymbol{\tau}_h\|_\infty \to 0$ *as* $h \to 0$*.*

Remark 3.2. With the Definition 3.1, all LMMs having at least kth-order truncation error are consistent of degree at least k. Moreover, since

$$\|\boldsymbol{\tau}_h\|_1 = \sum_{n=M}^{N} |(\boldsymbol{\tau}_h)_n| \leq N\|\boldsymbol{\tau}_h\|_\infty,$$

LMMs having at least second-order truncation are automatically consistent of degree 2 and thus strongly consistent.

Following from the classical truncation error analysis for LMMs, we have the algebraic criteria for the consistency.

LEMMA 3.3 (consistency). *A linear multistep method scheme for dynamics discovery is consistent provided that* $\rho(1) = 0$ *and* $\rho'(1) = \sigma(1)$*. Furthermore, it is consistent of degree* k *if it is order* k *in the sense of Definition* 2.3*, that is, if* $\rho(e^z) - z\sigma(e^z) = O(z^{k+1})$ *as* $z \to 0$*.*

3.4. Stability and the root condition for discovery. In this section we develop stability in a similar spirit as in section 2 but also introduce more refined notions of stability for convergence analysis.[1] For discovery, the main distinction from theory for solving the forward problem is that now we consider perturbations to

[1]It is interesting to note that some of our stability notions and root conditions are related to the A_∞ stability [36], often discussed in connection to the A-stability region, at infinity, of numerical integrators of known stiff problems, and relevant to solving differential algebraic equations (see [19], for example).

the recovered dynamics as opposed to the integrated states for the numerical solution of the differential equation. To begin we introduce a linear operator given by

$$(\hat{R}_h \hat{\boldsymbol{f}})_n := \sum_{m=m_0}^{M_0} \beta_m \hat{\boldsymbol{f}}_{n-m}, \quad n = M, M+1, \ldots, N. \tag{3.3}$$

Notice $(\hat{R}_h \hat{\boldsymbol{f}})_n$ arises from its forward counterpart (2.8) with the reduced $\boldsymbol{\beta}$ notation.

DEFINITION 3.4 (stability for dynamics discovery). *A linear M-step method for the dynamics discovery is called stable on $[a, b]$ provided there exists a constant $K < \infty$, not depending on N, such that, for any two grid functions $\boldsymbol{u}, \boldsymbol{v} \in \Gamma_h[a, b]$, we have*

$$\|\boldsymbol{u} - \boldsymbol{v}\|_\infty \leq K \left(\max_{i \in \mathcal{I}_M} |\boldsymbol{u}_i - \boldsymbol{v}_i| + \left\| \hat{R}_h(\boldsymbol{u} - \boldsymbol{v}) \right\|_\infty \right).$$

DEFINITION 3.5 (marginal stability for dynamics discovery). *A linear M-step method for the dynamics discovery is called marginally stable on $[a, b]$ provided that there exists a constant $K < \infty$, not depending on N, such that, for any two grid functions $\boldsymbol{u}, \boldsymbol{v} \in \Gamma_h[a, b]$, we have*

$$\|\boldsymbol{u} - \boldsymbol{v}\|_\infty \leq K \left(\max_{i \in \mathcal{I}_M} |\boldsymbol{u}_i - \boldsymbol{v}_i| + \left\| \hat{R}_h(\boldsymbol{u} - \boldsymbol{v}) \right\|_1 \right).$$

DEFINITION 3.6 (weak stability of degree $-k$ for dynamics discovery). *A linear M-step method for the dynamics discovery is called weakly stable of degree $-k$ for $k \geq 2$ on $[a, b]$ provided that there exists a constant $K < \infty$, not depending on N, such that, for any two grid functions $\boldsymbol{u}, \boldsymbol{v} \in \Gamma_h[a, b]$, we have*

$$\|\boldsymbol{u} - \boldsymbol{v}\|_\infty \leq K \left(N^{k-2} \max_{i \in \mathcal{I}_M} |\boldsymbol{u}_i - \boldsymbol{v}_i| + N^{k-1} \left\| \hat{R}_h(\boldsymbol{u} - \boldsymbol{v}) \right\|_\infty \right).$$

In all cases, the norm on the left-hand side is taken over the learned components $\{\mathbf{u}_n\}_{n \in \mathcal{I}}$ and $\{\mathbf{v}_n\}_{n \in \mathcal{I}}$. This convention is used in the rest of the paper. Note that we choose to use negative degree values so that more negative degrees correspond to weaker stability. Similar to the observation given in Remark 3.2, we see that weak stability of degree -2 follows from the marginal stability in Definition 3.5.

We would like to turn the property of stability into an algebraic condition as for the case of numerical solution to ODEs. For the forward problem, the algebraic root condition (Definition 2.6) serves this purpose; however, for the inverse problem, we require a more subtle treatment of the root condition to capture the nuances in stability for dynamics discovery.

DEFINITION 3.7 (strong root condition [1, 48, 13, 2]). *A polynomial satisfies the strong root condition provided the roots of the polynomial have magnitude less than* 1.

Likewise, we also generalize the above root conditions.

DEFINITION 3.8 (kth-multiplicity root condition). *A polynomial satisfies the root condition of degree $k \in \mathbb{N}$ provided the roots of the polynomial do not exceed magnitude* 1, *and those of magnitude* 1 *have multiplicity no larger than k.*

Remark 3.9. One may view the conventional (algebraic) root condition (Definition 2.6) and the strong root condition (Definition 3.8) as special cases of the kth-multiplicity root condition of Definition 3.8 with $k = 1$ and $k = 0$, respectively. The

strong root condition has been used in the numerical analysis, control theory, and linear recurrence relation literature for the study of the relative stability for LMM as time integrators and asymptotic properties associated with the linear recurrence relations [1, 48, 13, 2].

Naturally, we can see that the notions of stability for discovery for LMMs are tied to the bounds on the solutions to the linear recurrence equations determined by the coefficients $\boldsymbol{\beta}$. We now relate them to the root conditions. Notice that while the stability in Theorem 2.7 for numerical integration of the given dynamics is concerned with the first characteristic polynomial $\rho(r)$, the stability in Theorem 3.10 for the discovery of dynamics is concerned with the second characteristic polynomial $\sigma(r)$ defined by (2.3). More precisely, the root condition can be stated for a reduced second characteristic polynomial

$$\hat{\sigma}(r) = \sum_{m=m_0}^{M_0} \beta_m r^{M_0-m}. \tag{3.4}$$

Hence, we see a fundamental difference in the two stability notions. The dependence of stability on $\sigma(r)$ (or $\hat{\sigma}(r)$) might be unexpected as it has not appeared in the numerical differential equation literature. However, it is also not surprising given the inverse problem nature of using LMMs for dynamics discovery.

THEOREM 3.10 (stability for discovery). *A linear multistep method for discovery of dynamics is stable provided that the second characteristic polynomial $\sigma(r)$ or the reduced $\hat{\sigma}(r)$ satisfies the strong root condition in Definition* 3.7. *Likewise, an LMM for the discovery of dynamics is marginally stable provided that $\sigma(r)$ or $\hat{\sigma}(r)$ satisfies the algebraic root condition in Definition* 2.6. *Furthermore, an LMM for the discovery of dynamics is weakly stable of degree $-k$ (for $k \geq 2$) provided that $\sigma(r)$ or $\hat{\sigma}(r)$ satisfies the $(k-1)$th-multiplicity root condition in Definition* 3.8.

Proof. Let $\hat{\mathbf{e}} = \mathbf{u} - \mathbf{v}$, where $\mathbf{u}, \mathbf{v} \in \Gamma_h[0,T]$ are both generated by solving the LMM (3.2). By setting $\mathbf{r} = \hat{R}_h(\mathbf{e})$ with the operator $\hat{R}_h$ defined in (3.3), we have

$$\sum_{m=m_0}^{M_0} \beta_m \hat{\mathbf{e}}_{n-m} = \mathbf{r}_n, \quad n = M, M+1, \ldots, N.$$

By standard recurrence and linear algebra theory [1, 15], the difference $\hat{\mathbf{e}}$ can be determined by the companion matrix of the above recurrence relation, denoted by $\mathcal{Z}$. This matrix is an $(M_0 - m_0) \times (M_0 - m_0)$ matrix with its first row given by $-(\frac{\beta_{m_0+1}}{\beta_{m_0}}, \frac{\beta_{m_0+2}}{\beta_{m_0}}, \ldots, \frac{\beta_{M_0}}{\beta_{m_0}})$, and the rest of the rows are of the form $(\mathbf{I}, \mathbf{0})$, where $\mathbf{I}$ is the identity matrix of size $(M_0 - m_0 - 1) \times (M_0 - m_0 - 1)$ and $\mathbf{0}$ is the zero column vector in $\mathbb{R}^{M_0-m_0-1}$. The matrix $\mathcal{Z}$ is associated with a characteristic polynomial given by $\hat{\sigma}(r)$ that shares the same set of roots as that of $\sigma(r)$, except a possible root at 0.

To consider the propagation of the difference $\hat{\mathbf{e}}$, we form the matrix $\mathbf{E}_n = Z\mathbf{E}_{n-1} + \mathbf{R}_n$, where $\mathbf{E}_n \in \mathbb{R}^{(M_0-m_0)\times d}$ has its rows given by $\{\hat{\mathbf{e}}_{n-k}\}_{0 \leq k < M_0-m_0}$, and $\mathbf{R}_n \in \mathbb{R}^{(M_0-m_0)\times d}$ has its first row given by the vector $\beta_{m_0}^{-1}\mathbf{r}_n$, and all subsequent rows by zeros.

Then

$$\mathbf{E}_n = Z^{n-M+1}\mathbf{E}_{M-1} + \sum_{k=M}^{n} Z^{n-k}\mathbf{R}_k,$$

where $\mathbf{E}_{M-1}$ is given by the initial data $\{\hat{\mathbf{e}}_k\}_{k\in\mathcal{I}_M}$. Thus, stability is equivalent to

$$\max_{M\le n\le N}\|\mathbf{E}_n\|_\infty \le K(\|\mathbf{E}_{M-1}\|_\infty + \max_{M\le n\le N}\|\mathbf{R}_n\|_\infty),$$

which is implied by

$$\sum_{n=1}^{N_M}\|\mathcal{Z}^n\|_\infty \le K^* < \infty,$$

or equivalently the strong root condition. Meanwhile, marginal stability is equivalent to

$$\max_{M\le n\le N}\|\mathbf{E}_n\|_\infty \le K(\|\mathbf{E}_{M-1}\|_\infty + \sum_{M\le n\le N}\|\mathbf{R}_n\|_\infty),$$

which is implied by

$$\max_{1\le n\le N_M}\|\mathcal{Z}^n\|_\infty \le K^* < \infty.$$

We thus only need the algebraic root condition.

Likewise, we can argue that weak stability of degree $-k$ is implied by

$$\max_{1\le n\le N_M}\|\mathcal{Z}^n\|_\infty \le K^* N^{k-2} < \infty, \quad \text{and} \quad \sum_{n=1}^{N_M}\|\mathcal{Z}^n\|_\infty \le K^* N^{k-1},$$

which is equivalent to the $(k-1)$th multiplicity root condition. ∎

3.5. Error analysis and convergence. In this section, we use the truncation error to study the error for discovery, including defining convergence and the order of approximation of LMM schemes for discovery.

DEFINITION 3.11 (convergence and order of approximation for discovery). *Consider the initial value problem* (2.1) *discretized by an M-step LMM given by* (3.1). *Let $\boldsymbol{f}, \hat{\boldsymbol{f}} \in \Gamma_h[a,b]$, where $\boldsymbol{f}$ is the exact grid function on the $N+1$ grid points $\{\boldsymbol{f}_n = f(x(t_n))\}$ and $\hat{\boldsymbol{f}}$ the approximation solved from* (3.1). *The LMM is convergent for dynamics discovery if $\|\boldsymbol{f}-\hat{\boldsymbol{f}}\|_\infty \to 0$ as $h\to 0$ whenever $\max_{i\in\mathcal{I}_M}|\boldsymbol{f}_i - \hat{\boldsymbol{f}}_i| \to 0$. Moreover, if $\|\boldsymbol{f}-\hat{\boldsymbol{f}}\|_\infty = ch^p + O(h^{p+1})$ for some constant c, then p is called the convergence order, or alternatively, the order of approximation for dynamics discovery.*

Using the introduced notions of consistency and stability, we now present convergence theorems for dynamics discovery.

THEOREM 3.12 (convergence theorems for discovery I). *Consider the dynamical system* (2.1) *discretized by an M-step LMM given by* (3.1). *Let $\boldsymbol{f}, \hat{\boldsymbol{f}} \in \Gamma_h[a,b]$, where $\boldsymbol{f}$ is the exact grid function on the $N+1$ grid points $\{\boldsymbol{f}_n = f(x(t_n))\}$ and $\hat{\boldsymbol{f}}$ the approximation solved from* (3.1). *Then,*

$$B(\hat{\boldsymbol{f}} - \boldsymbol{f}) = \boldsymbol{\tau}_h + \boldsymbol{g}_h, \tag{3.5}$$

where $\boldsymbol{\tau}_h$ is the local truncation error of the scheme, and $\boldsymbol{g}_h = (\boldsymbol{g}_M, \boldsymbol{g}_{M+1}, \dots, \boldsymbol{g}_N)$ is given by

$$(\boldsymbol{g}_h)_n = \begin{cases} \displaystyle\sum_{\substack{m\ge n-M_0\\ m\in\mathcal{I}_M}} \beta_{n-m}(\boldsymbol{f}_m - \hat{\boldsymbol{f}}_m) & \text{if } n \in M_0 + \mathcal{I}_M, \\ 0 & \text{if } n \in \mathcal{I}\setminus\{M_0+\mathcal{I}_M\}. \end{cases}$$

Moreover, in the senses of consistency outlined in Definition 3.1 *and stability in Definitions* 3.4–3.6*, if an LMM is consistent and stable, or strongly consistent and marginally stable, then it is convergent for dynamics discovery in the sense of Definition* 3.11*. Furthermore, if it is consistent of degree k and weakly stable of degree $-k$, then provided that $N^{k-2}\max_{i\in\mathcal{I}_M}|\boldsymbol{f}_i-\hat{\boldsymbol{f}}_i|\to 0$ as $h\to 0$, we also have $\|\boldsymbol{f}-\hat{\boldsymbol{f}}\|_\infty\to 0$.*

Proof. By (3.1) and the truncation error defined in (2.9), we have

$$\sum_{m=0}^{M}\frac{1}{h}\alpha_m\mathbf{x}_{n-m}-\sum_{m=m_0}^{M}\beta_m\hat{\boldsymbol{f}}_{n-m}=\mathbf{0},$$
$$\sum_{m=0}^{M}\frac{1}{h}\alpha_m\mathbf{x}_{n-m}-\sum_{m=m_0}^{M}\beta_m\boldsymbol{f}_{n-m}=(\boldsymbol{\tau}_h)_n.$$

Subtracting the equations, we observe

$$\sum_{m=m_0}^{M}\beta_m(\hat{\boldsymbol{f}}_{n-m}-\boldsymbol{f}_{n-m})=(\boldsymbol{\tau}_h)_n,\quad n=M,M+1,\dots,N, \tag{3.6}$$

or equivalently $B(\boldsymbol{f}-\hat{\boldsymbol{f}})=\boldsymbol{\tau}_h+\boldsymbol{g}_h$, where $\hat{\boldsymbol{f}}-\boldsymbol{f}$ on the left side refers to those components indexed in $\mathcal{I}$.

Now, by the definitions of stability given in Definitions 3.4 and 3.5, there exists a constant $K_W<\infty$ independent of h, for h sufficiently small, such that

$$\left\|\boldsymbol{f}-\hat{\boldsymbol{f}}\right\|_\infty\le K_W\left(\max_{i\in\mathcal{I}_M}\left|\boldsymbol{f}_i-\hat{\boldsymbol{f}}_i\right|+\|\boldsymbol{\tau}_h\|_W\right),$$

where $W=\infty$ or $W=1$, if the LMM is stable or marginally stable, respectively. Thus, by Definition 3.1 on consistency and strong consistency, we have $\|\boldsymbol{f}-\hat{\boldsymbol{f}}\|_\infty\to 0$ as $h\to 0$. Likewise, if the LMM is weakly stable of degree $-k$, then

$$\left\|\boldsymbol{f}-\hat{\boldsymbol{f}}\right\|_\infty\le K\left(N^{k-2}\max_{i\in\mathcal{I}_M}\left|\boldsymbol{f}_i-\hat{\boldsymbol{f}}_i\right|+N^{k-1}\left\|\boldsymbol{\tau}_h\right\|_\infty\right).$$

By the definition of the consistency of degree k, together with the assumption on the initial data that $N^{k-2}\max_{i\in\mathcal{I}_M}|\boldsymbol{f}_i-\hat{\boldsymbol{f}}_i|\to 0$, convergence also follows. ∎

Theorem 3.12 states that for LMM based dynamics discovery, convergence follows from both stability and consistency, as in the case of LMM-based time integration. Equation (3.5) shows the interplay between the stability aspect of solving the system, manifested in B^{-1}, and the consistency component of truncation error, $\boldsymbol{\tau}_h$, from discretization of the differential equation.

We note that Theorem 3.12 contains only sufficient conditions for convergence. There are examples of LMMs that are consistent and marginally stable, but not strongly consistent nor stable, which may still be convergent for dynamics discovery. An example is the LMM with $\rho(r)=r^2-1$ and $\sigma(r)=(r+1)/2$; convergence for this LMM can be checked using calculations similar to that presented in the proof of Corollary 4.3. Nevertheless, in the spirit of the Dahlquist equivalence theorem, we also have the following result establishing consistency and some form of stability from convergence.

THEOREM 3.13 (convergence theorems for discovery II). *Consider the dynamical system* (2.1) *discretized by an M-step LMM given by* (3.1). *If the LMM is convergent for dynamics discovery in the sense of Definition* 3.11, *then it is consistent and marginally stable in the senses of Definitions* 3.1 *and* 3.5.

Proof. The proof is similar to its classical counterpart. Consider the special cases of ODE $\frac{d}{dt}\mathbf{x}(t) = 0$ and $\frac{d}{dt}\mathbf{x}(t) = 1$, respectively, with $\mathbf{x}(a) = 0$. If the LMM is convergent in the sense of Definition 3.11, then the dynamical system (3.1) leads to a linear recurrence relation with constants $\rho(1)$ or $\rho'(1)$, respectively, serving as inhomogeneous terms on the right-hand side. Since the LMM is convergent, the learned dynamics approach 0 or 1, respectively. Thus, as $h \to 0$, we get $\rho(1) = 0$ and $\rho'(1) = \sigma(1)$. Consequentially, the consistency conditions from Lemma 3.3 are satisfied. Using the theory on the linear recurrence relations given in the proof of Theorem 3.12, in order for $\|\boldsymbol{f} - \hat{\boldsymbol{f}}\|_\infty \to 0$ as $h \to 0$ whenever $\max_{i\in\mathcal{I}_M}|\boldsymbol{f}_i - \hat{\boldsymbol{f}}_i| \to 0$, there must exist some constant $0 < K < \infty$ such that $\max_{1\le n\le N_M}\|\mathcal{Z}^n\| < K$ as $N_M \nearrow \infty$. For this bound to exist, the root condition must be satisfied, and hence the method must be marginal stable. ∎

As seen from the proof of Theorem 3.12, under some assumptions on the initial dynamics, we immediately get the order of convergence for LMM-based dynamics discovery.

THEOREM 3.14 (order of convergence). *If an LMM has truncation error of order k with $k \ge 1$, i.e., $\|\boldsymbol{\tau}_h\|_\infty = O(h^k)$, as in Definition* 2.3, *then, as $h \to 0$, we have $\|\boldsymbol{f} - \hat{\boldsymbol{f}}\|_\infty = O(h^k)$ if the LMM is stable and $\max_{i\in\mathcal{I}_M}|\boldsymbol{f}_i - \hat{\boldsymbol{f}}_i| = O(h^k)$. Moreover, provided that $\max_{i\in\mathcal{I}_M}|\boldsymbol{f}_i - \hat{\boldsymbol{f}}_i| = O(h^{k-1})$, we have $\|\boldsymbol{f} - \hat{\boldsymbol{f}}\|_\infty \le Ch^{k-1}$ if it is marginally stable or $\|\boldsymbol{f} - \hat{\boldsymbol{f}}\|_\infty \le Ch^{k+1-s}$ if it is weakly stable of degree $-s$ with $k \ge \max\{s, 2\}$.*

Remark 3.15. The different notions of stability affect the order of convergence for dynamics discovery. These refinements motivate accompanying definitions for the degree of consistency in Definition 3.1, whereas traditionally the order of error matches with the order of convergence (see Definition 2.3). For dynamics discovery, the order of convergence and degree of consistency matches for strongly stable schemes. While this might not hold generically for marginally or weakly stable LMMs, resulting in possible lower order of convergence than the degree of consistency, we show later that, for some cases such as AM-1, the same order can still be maintained.

4. Application to AB, AM, and BDF.

We now apply the general theorem on LMM for the dynamics discovery to three popular special classes of methods—AB, AM, and BDF.

4.1. Consistency of AB, AM, and BDF.

It is well-known that the Adams family schemes and BDF are consistent as time integrators. Specifically, AB-M and BDF-M have order of error M, while AM-M has order of error $M+1$. As a result, these three classes of LMM methods remain consistent for dynamics discovery. Moreover, as a consequence of the order of error, AB-M and BDF-M are consistent of degree M, and the AM-M schemes are consistent of degree $M + 1$, as noted in Remark 3.2. Indeed, the latter fact is crucial to the convergence of AM-1.

THEOREM 4.1 (consistency of AB, AM, and BDF for dynamics discovery). *The AB-M, AM-M, and BDF-M schemes are all consistent for dynamics discovery. Furthermore, AM-1 is consistent of degree* 2 *and thus strongly consistent.*

4.2. Stability and convergence of AB, AM, and BDF.

THEOREM 4.2. *With the notions of stability defined in Definitions* 3.5 *and* 3.4,
1. *BDF-M for all* $M \geq 1$, *AB-M for* $1 \leq M \leq 6$, *and AM-*0 *are stable;*
2. *AM-*1 *is marginally stable and thus weakly stable of degree* -2*;*
3. *AB-M for* $7 \leq M \leq 10$ *and AM-M for* $M \geq 2$ *are unstable.*

The proof of Theorem 4.2 is given in section 4.3.

COROLLARY 4.3. *BDF-M for all* $M \geq 1$ *are convergent, with convergence order* M. *AB-M for* $1 \leq M \leq 6$ *are convergent, with convergence order* M. *AM-*0 *is convergent with first-order convergence. AM-*1 *is convergent with second-order convergence if we have second-order error on the initial data.*

Proof. The conclusions of Corollary 4.3 on the convergence of LMM schemes under consideration follow immediately from the application of Theorems 4.1, 4.2, and 3.14. The order of convergence follows, with the exception of AM-1. Indeed, a direct application would imply only first-order convergence due to its degree-1 marginal stability. However, we note that in this special case, the recurrence relation (3.6) is given by $\hat{\boldsymbol{f}}_n - \boldsymbol{f}_n = -(\hat{\boldsymbol{f}}_{n-1} - \boldsymbol{f}_{n-1}) + 2(\boldsymbol{\tau}_h)_n$. Using the error expansion given in Definition 2.3, the leading order of $\hat{\boldsymbol{f}}_n - \boldsymbol{f}_n$ of the form

$$(-1)^{n-j}(\hat{\boldsymbol{f}}_0 - \boldsymbol{f}_0)h^2 + \sum_{j=1}^{n}(-1)^{n-j}\mathbf{e}(t_j)h^2$$

$$\approx O(h^2) + \left\{ \begin{array}{ll} \displaystyle\sum_{j=1}^{k} \mathbf{e}'(t_{2j})h^3 & \text{if } n = 2k, \\ \displaystyle\sum_{j=1}^{k} \mathbf{e}'(t_{2j+1})h^3 + (-1)^{n-1}\mathbf{e}(t_1)h^2 & \text{if } n = 2k+1 \end{array} \right\} \approx O(h^2),$$

where $\mathbf{e}(t) = x^{(p+1)}(t)$ is assumed to be a smooth function depending on the solution $x = x(t)$ of the exact dynamics. Therefore, given second-order error in the initial data, AM-1 has second-order convergence even though it is not a strongly stable method. ∎

Remark 4.4. As demonstrated in the proof of the above Corollary, we see that if all the roots of $\hat{\sigma}$ on the unit circle are also the roots of 1, some refined notions of the consistency (such as requiring the difference of truncation errors at two consecutive times steps being of a higher order) can be developed to utilize the error cancellation to maintain convergence and error order. We note also that the finite range of instability with respect to the order M for the AB scheme is due to the limitation of using brute force calculations, but we conjecture that the scheme is unstable for all $M \geq 7$. Interestingly, that $M = 6$ is a threshold for stability of the polynomial echoes the stability criterion for the forward problem BDF [22], for which $M = 6$ is also the largest known order method that is stable. Explicit numerical calculation or Routh arrays (see [13]) are used to show this fact [22, 10, 14]. Schur polynomials have since been used [9] to show a generalized stability argument for $M \geq 13$ [14]. We leave open a generalized stability result for $M \geq 7$ using the polynomial roots, but we have validated numerically the instability for $7 \leq M \leq 20$.

4.3. Verification of root conditions for AB, AM, and BDF. We now verify, for the three classes of LMMs, that the root condition holds for cases stated in Theorem 4.2.

TABLE 1
Largest magnitude roots.

Step M	1	2	3	4	5
AB	–	0.3333	0.4663	0.6338	0.8075
AM	1.0000	1.7165	2.3658	2.9775	3.5639
Step	6	7	8	9	10
AB	0.9829	1.1587	1.3345	1.5100	1.6852
AM	4.1312	4.6851	5.2267	5.7586	6.2820

We begin by calculating the roots of the second characteristic polynomial associated with AB and AM since $\sigma(r) = \hat{\sigma}(r)$ in both cases. We first present some results for AB-M and AM-M with $1 \leq M \leq 10$ as computational evidence (with exact symbolic computation). We have also numerically validated instability for AB-M for $11 \leq M \leq 20$ as well and expect instability to persist for all $M \geq 7$. However, there is no theoretical proof so far. For AM-M, a general instability result for $M \geq 2$ is proved in Lemma 4.7.

Fix $M \in \mathbb{N}$ and $\tilde{\Lambda} \in \{\Lambda_0, \Lambda_1\}$, where we recall from section 2 that $\Lambda_0 = \{-M, \ldots, 0\}$ and $\Lambda_1 = \{-M, -M+1, \ldots, -1\}$. Exchanging the integral and the summand in the formula for the Lagrange interpolating polynomial, one can observe that finding the roots of the second characteristic polynomial is equivalent to choosing $r \in \mathbb{C}$ satisfying a mean-zero equation. That is, for $\ell_x^h(u; \tilde{\Lambda})$ defined in (2.6), we have

$$\sum_{x \in \tilde{\Lambda}} \int_0^1 \ell_x^h(u; \tilde{\Lambda}) r^x du \iff \int_0^1 \sum_{x \in \tilde{\Lambda}} \ell_x^h(u; \tilde{\Lambda}) r^x du = 0. \tag{4.1}$$

As we see in Table 1, which is computed symbolically using Mathematica, the profile of the roots of the characteristic polynomial associated with the different schemes varies significantly. Equation (4.1) and the data in Table 1 immediately lead to the following lemma.

LEMMA 4.5. *Fix $\tilde{\Lambda} \in \{\Lambda_0, \Lambda_1\}$, and let $\ell_x^h(u; \tilde{\Lambda})$ be the Lagrange interpolating polynomial defined in* (2.6)*. Then, we can characterize the roots $r \in \mathbb{C}$ of the second characteristic polynomial as the solution to the equation*

$$\int_0^1 \sum_{x \in \tilde{\Lambda}} \ell_x^h(u; \tilde{\Lambda}) r^x du = 0. \tag{4.2}$$

Moreover, for AM-M with $\tilde{\Lambda} = \Lambda_0$, we have

1. *$M = 1$, then the single root satisfies $|r| = 1$,*
2. *$2 \leq M \leq 10$, there exists at least one root r with $|r| > 1$,*

and for AB-M with $\tilde{\Lambda} = \Lambda_1$, we have

1. *$1 \leq M \leq 6$, then $|r| < 1$,*
2. *$7 \leq M \leq 10$, there exists at least one root r with $|r| > 1$.*

Let us state some useful properties of the second characteristic polynomial $\sigma(r)$ associated with the AM methods and the corresponding coefficients of its B matrix.

LEMMA 4.6. *For $M \geq 2$, the coefficients $\{\beta_m\}_0^M$ of the AM method have the following properties:*

1. *$\beta_1 > \beta_0 > 0$,*
2. *$\operatorname{sign}(\beta_{m+1}) = -\operatorname{sign}(\beta_m)$, $1 \leq m \leq M-1$, and*
3. *$\beta_0 > |\beta_M|$.*

Proof. Fix $M \in \mathbb{N}$ with $M \geq 2$. The M-step AM scheme has coefficients

$$\beta_m = \frac{(-1)^m}{m!(M-m)!} \int_0^1 \prod_{\substack{i=0 \\ i \neq m}}^{M} (u+i-1)\, du \tag{4.3}$$

for $m = 0, 1, \ldots, M$. The coefficients β_0 and β_1 are given by

$$\beta_0 = \frac{1}{M!} \int_0^1 \prod_{i=0}^{M-1} (u+i) du \quad \text{and} \quad \beta_1 = \frac{1}{(M-1)!} \int_0^1 (1-u) \prod_{i=1}^{M-1} (u+i) du.$$

Certainly, $\beta_0 > 0$. Notice

$$\beta_1 > \beta_0 \iff \frac{M}{M+1} \int_0^1 \prod_{i=1}^{M-1} (u+i) du > \int_0^1 \prod_{i=0}^{M-1} (u+i) du. \tag{4.4}$$

We prove (4.4) by induction. As the base case, $M = 2$. For $M = 2$, we have $\beta_1 = 8/12 > \beta_0 = 5/12$. Now assume (4.4) holds up to some arbitrary $M \in \mathbb{N}$ with $M > 2$. We will show the result for $M + 1$.

(4.5)

$$\frac{M+1}{M+2} \int_0^1 \prod_{i=1}^{M} (u+i) du = \frac{M+1}{M+2} \int_0^1 \left(u \prod_{i=1}^{M-1} (u+i) + M \prod_{i=1}^{M-1} (u+i) \right) du$$

$$\overset{(4.4)}{>} \frac{M+1}{M+2} \left(\int_0^1 \prod_{i=0}^{M-1} (u+i) + \frac{M(M+1)}{M} \prod_{i=0}^{M-1} (u+i) du \right) \tag{4.6}$$

$$= \frac{(M+1)(M+2)}{(M+2)} \int_0^1 \prod_{i=0}^{M-1} (u+i) du \tag{4.7}$$

$$> (M+1) \int_0^1 \frac{u+M}{M+1} \prod_{i=0}^{M-1} (u+i) du = \int_0^1 \prod_{i=0}^{M} (u+i) du, \tag{4.8}$$

as desired. Note we used the inductive hypothesis on the second term in (4.6). The proof by induction showing for $M \geq 2$, $\beta_1 > \beta_0$ is complete. To prove part 2, note that the relation of signs between coefficients follows from the sign of the Lagrange basis polynomials in the integrand of the coefficients. For $m \in \{2, 3, \ldots, M\}$, the integrand of (4.3) are of the same sign, and therefore the sign of β_m depends only on the multiplier $(-1)^m$. Hence part 2 of Lemma 4.6 follows.

Finally, for part 3, we note that

$$|\beta_M| = \frac{1}{M!} \int_0^1 (1-u) \prod_{i=0}^{M-2} (u+i)\, du < \frac{1}{M!} \int_0^1 \prod_{i=0}^{M-2} (u+i)\, du$$
$$< \frac{1}{M!} \int_0^1 (u+M-1) \prod_{i=0}^{M-2} (u+i)\, du = \beta_0.$$

This completes the proof. ∎

LEMMA 4.7 (general instability of AM $M \geq 2$). *The linear multistep method formed by the AM scheme for $M \geq 2$ does not satisfy the root condition.*

Proof. Fix $M \geq 2$ and consider the second characteristic polynomial associated with the AM scheme. We write it as $\sigma(r) = \sum \beta_m r^{M-m}$. From Lemma 4.6, $\beta_1/\beta_0 > 1$. Moreover, by construction of the AM method, $(-1)^m \beta_m < 0$ for $m \geq 2$.

For $r > 0$ sufficiently large.

$$(-1)^M \sigma(-r) = (-1)^{2M} r^M \left[\beta_0 - \beta_1/r + \sum_{m=2}^{M} (-1)^m \beta_m r^{-m} \right]. \tag{4.9}$$

Taking the limit as $r \to +\infty$, we see that $(-1)^M \sigma(-\infty) = \infty$ since $\beta_0 > 0$. Meanwhile,

$$(-1)^M \sigma(-\beta_1/\beta_0) = \sum_{m \geq 2} (-1)^{-m} \beta_m (\beta_1/\beta_0)^{M-m} < 0.$$

Hence, it follows from the intermediate value theorem that there is at least one root of $\sigma(r)$ that is real in $(-\infty, -\beta_1/\beta_0) \subset (-\infty, -1)$, violating the root condition. The result thus follows. ∎

THEOREM 4.8 (root condition of AB, AM, BDF). *The strong root condition for discovery is satisfied by BDF-M for all $M \in \mathbb{N}$, the AB-M scheme for $1 \leq M \leq 6$, and AM-M for $M = 0$. The algebraic root condition, or the kth root condition with $k = 1$, is satisfied for AM-M with $M = 1$. On the other hand, the root condition is not satisfied for the AB-M scheme with $7 \leq M \leq 10$ or the AM-M scheme with $M \geq 2$.*

Proof. The case of AM-0 is trivial. Lemma 4.5 implies the results of Theorem 4.8 for AB-M with $1 \leq M \leq 10$ and for AM-M with $1 \leq M \leq 10$. Furthermore, by Lemma 4.7, the AM-M scheme for $M \geq 2$ violates the root condition and hence is unstable. Finally, BDF-M has $\sigma(r) = r^{M-1}$ and $\hat{\sigma}(r) = 1$ for all $M \geq 1$. Hence, the root condition is always satisfied for the BDF scheme for arbitrary $M \geq 1$. As a result, AM-0, identical to BDF-1, satisfies the root condition as well. ∎

Finally, Theorem 4.2 follows directly from Theorems 4.8 and 3.10.

4.4. Discussions on the effect of initial conditions. The theory developed so far is under the assumption that some initial data of the dynamics are provided, which leads to learning the approximated dynamics at later times. One may consider a situation where the some terminal data are given instead. In such cases, the approximate dynamics would be solved backward in time, yielding a modified system of equations. It is not hard to check that the stability would become dependent on a modified second characteristic polynomial whose roots are the reciprocals of those of $\hat{\sigma}$. Naturally, it is of interest to check root conditions for the three classes of LMMs as well. For BDF, we clearly see the strong root condition holds as $\hat{\sigma}(r) = 1$. For AM-0 and AB-1, the same also hold. Likewise for AM-1, the root condition but not the strong root condition remains true. For AM-M with $M \geq 2$, part 3 of Lemma 4.6 implies that the product of the roots of $\hat{\sigma}(r) = \sigma(r)$ is less than one. Therefore, there might be at least one root of the modified second characteristic polynomial outside the unit disc, and hence instability for these methods is again expected. Interestingly, unlike in the case with initial data where there is not yet rigorous theory but only computational results for the AB methods, one can prove rigorously in the next lemma a result of instability for the backwards-in-time AB-M, $M \geq 2$, via a similar argument as in part 3 of Lemma 4.6.

LEMMA 4.9. *For $M \geq 2$, the coefficients $\{\beta_m\}_0^M$ of AB-M satisfy $\beta_0 = 0$ and $\beta_1 > |\beta_M|$.*

Proof. $\beta_0 = 0$ is true by construction. For $M \geq 2$, we have

$$\beta_m = \frac{(-1)^{m+1}}{(m-1)!(M-m)!} \int_0^1 \prod_{\substack{i=1 \\ i \neq m}}^{M} (u+i-1)\, du \tag{4.10}$$

for $m = 1, 2, \ldots, M+1$. The coefficients β_1 and β_M satisfy

$$|\beta_M| = \frac{1}{(M-1)!} \int_0^1 \prod_{i=1}^{M-1} (u+i-1) du < \frac{1}{(M-1)!} \int_0^1 \prod_{i=2}^{M} (u+i-1) du = \beta_1,$$

which completes the proof of the lemma. ∎

From the above, we see that root conditions do not hold for the modified second characteristic polynomial associated with AB-M with $M \geq 2$, so that instability would occur when terminal data are supplied. In practice, it is often the case that such initial dynamics are represented by neural networks as part of the unknown as well. Thus, the stability in such cases is worthy of further investigation, particularly in conjunction with the approximation properties of the neural networks to be employed. Clearly, the successful runs using neural networks in Figure 1 have good correspondence with those schemes enjoying some stability properties in one or both types of initial/terminal data.

5. Long-time dynamics discovery. In this section, we consider the problem of discovering dynamics of (2.1) over a variable interval $(0, T)$, with terminal time $1 \ll T \to \infty$, and a fixed mesh h. Notice by increasing T we increase the number of grid points $N = T/h$; hence we hope to relate our previous studies with variable mesh and fixed domain to this setting. For the numerical analysis of time integration, this study is reminiscent of that of asymptotic stability, which is often treated via the study of linear dynamics [15, 33, 2].

By rescaling time, $\tilde{t} = t/T$, where $0 \leq \tilde{t} \leq 1$, and defining $\tilde{x}(\tilde{t}) = x(t)$, we have via change of variables that the scaled dynamics $\tilde{f}$ may be related to that of the original variables by

$$\frac{d}{d\tilde{t}} \tilde{x}(\tilde{t}) = T \frac{d}{dt} x(t) = T f(x(t)) = T f(\tilde{x}(\tilde{t})).$$

Then, if we define $\tilde{f}(\tilde{x}(\tilde{t})) = T f(\tilde{x}(\tilde{t})) = T f(x(t))$, the rescaled differential equation becomes

$$\frac{d}{d\tilde{t}} \tilde{x}(\tilde{t}) = \tilde{f}(\tilde{x}(t)), \quad 0 \leq \tilde{t} \leq 1, \quad \tilde{x}(0) = x(0) = x_0. \tag{5.1}$$

Now, consider applying the LMM scheme to $\tilde{x}$ using the transformed model problem (5.1) with a step size $\tilde{h} = 1/N$. Under this rescaling of time, one can check directly that the leading truncation error term of an LMM of order p in the sense of Definitions 2.2 and 2.3 is

$$C_{p+1} \tilde{h}^p \frac{d^{p+1}}{d\tilde{t}^{p+1}} \tilde{x}(\tilde{t}) = C_{p+1} \tilde{h}^p T^{p+1} \frac{d^{p+1}}{dt^{p+1}} x(t) = C_{p+1} T h^p \frac{d^{p+1}}{dt^{p+1}} x(t). \tag{5.2}$$

In light of (5.2), we can see that the truncation error of the discovered dynamics of (2.1) in the original time scale is a multiple of the truncation error of the rescaled model (5.1) by the factor T^{-1}. Meanwhile, from the analysis of section 3.4, the stability bound in Definition 3.4 is only directly dependent on $\sigma(r)$, not the specific time domain size for $T > 1$.

Using these observations of the effects on consistency and stability, we can deduce the behavior of an LMM in the long-time regime. For a strongly stable pth-order LMM, the global error behaves like $O\left(T^{-1}Th^p\right) = O(h^p)$ provided that $\max_{t\in(0,T)} |x^{(p+1)}(t)|$ remains uniformly bounded as T increases. Hence, we may view strongly stable LMMs as A-stable, in the case of dynamics discovery, for fixed h as $T \to \infty$. This can be seen as another difference with the case of the forward problem of time integration, where the order of A-stable LMMs is known to be limited by 2 due to the celebrated Dahlquist barrier theorems [12, 15, 33, 2]. On the other hand, for unstable methods, the exponential growth in N of the inverse matrix B^{-1} dominates over any gain in accuracy from consistency. Thus, lack of stability leads to an exponentially increasing error as T grows linearly.

As a special example, the marginally stable AM-1 is not stable for dynamics discovery, but as stated in the Corollary 4.3 and the derivation in its proof, we can use the rescaling to get the global error in the form $O(T^{-1}Th^2) = O(h^2)$, Thus, we expect AM-1, for a fixed h, to have a constant error as T increases, which is supported by numerical experiments presented in the next section.

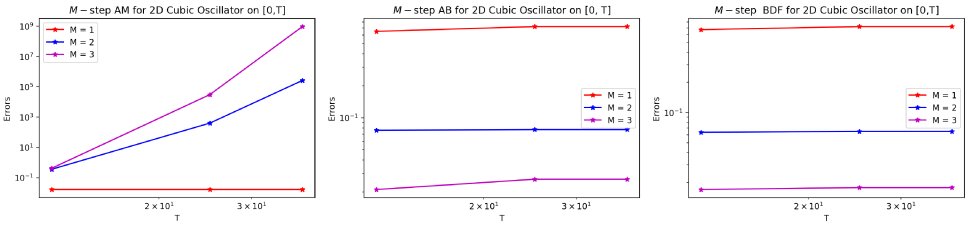
To recap, from the analysis in this section, for dynamics discovery, BDFs enjoy asymptotic stability for a fixed time step size h as T increases. The same holds for AB-M, at least for a small value of M that enjoys the stability as $h \to 0$ for a given terminal time. While this also holds for AM-1, it does not hold for AM-M with $M \geq 2$. As shown in Figure 2, the errors from AB and BDF remain fixed across various values of T, while the AM methods yield exponential growth of error in T for $M \geq 2$.

6. Numerical experiments. In this section, we provide numerical solutions to the linear systems associated with each of the studied multistep methods and show numerical evidence consistent with the theoretical findings. We limit ourselves to the idealized setting of numerically exact states considered for the theoretical analysis and to low dimensional dynamic systems for the sake of illustration and benchmarking. In addition, we also take the initial data for the dynamics to be exact. For a model problem, we consider the 2D cubic system, a nonlinearly damped oscillator, specified as in [41, 6].

$$\begin{cases} \dot{x}_1 = -0.1\ x_1^3 + 2.0\ x_2^3, \\ \dot{x}_2 = -2.0\ x_1^3 - 0.1\ x_2^3, \\ x_1(0) = 2,\ x_2(0) = 0. \end{cases} \tag{6.1}$$

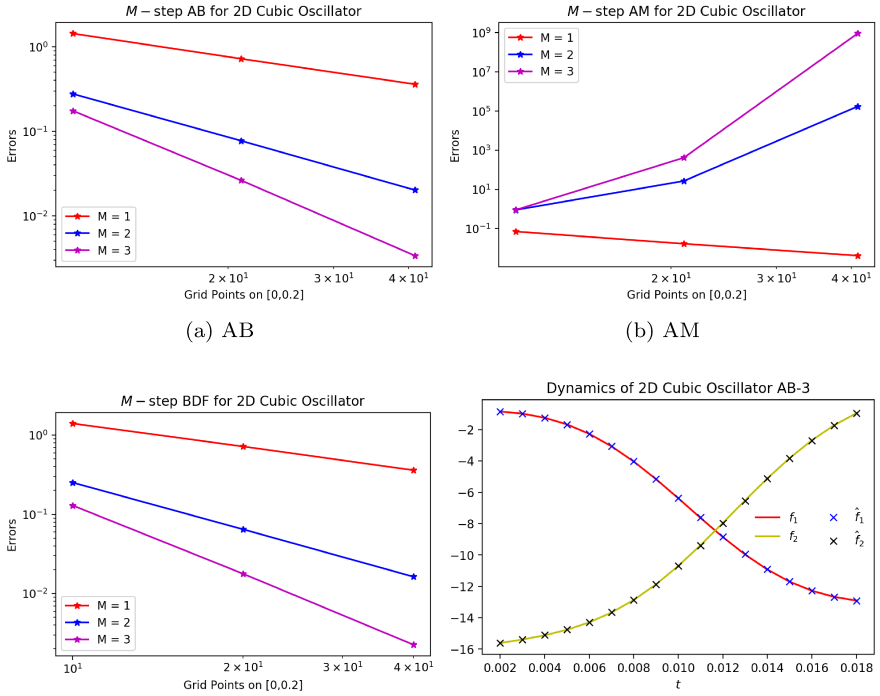
6.1. Fixed time domain. First we study the methods on a fixed time domain, $t \in [0, 0.2]$, with varying time step. We show in Figure 3 the results from the Adams family and BDF methods.

The exact states and dynamics are computed by numerically integrating (6.1) on a very refined mesh. The errors of the discovered dynamics in the $\ell^\infty-$norm are shown in Figure 3 for a different M against a different number of grid points. In addition, Figure 3(d) shows a segment of the approximated dynamics captured over the interval versus the true dynamics using a stable and convergent method (AB-3) when $h = 0.01$.



(a) AM (b) AB (c) BDF

FIG. 2. *Long-time errors for discovery of 2D cubic system* (6.1).



(a) AB (b) AM

(c) BDF (d) Captured dynamics with AB-3

FIG. 3. *Numerical results of the three types of schemes on the 2D cubic system* (6.1) *on the unit time interval for different choices of* M *and* N.

In this figure, the dotted and dashed lines represent the true dynamics in the first and second coordinates, i.e., $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$, respectively. The crosses and asterisks denote the learned dynamics in the first and second coordinates, i.e., $\hat{\boldsymbol{f}}_1$ and $\hat{\boldsymbol{f}}_2$, respectively. The method is able to capture the twist and intersection of the two coordinates. Clearly, the numerical results support the theoretical findings of this paper.

6.2. Long-time behavior. Here, we consider the problem of discovering dynamics over a changing domain $[0, T]$, for $T \gg 1$, with fixed mesh size h. In Figure 2, we discover the dynamics of the 2D cubic system over specified ranges of T ($T = 12.5, 25, 37.5$). In Figures 2(a), 2(b), and 2(c), we use $h = 0.01$ to first generate data over $[0, 50]$ and then select the slice of data matching the Ts. AM-M clearly suffers from the exponential error growth when $M \geq 2$, while it has a constant error

TABLE 2
LMMs: Similarities and differences for integrating and learning dynamics.

Task	Integrating dynamics	Learning dynamics
Goal	Given $\boldsymbol{f} = \boldsymbol{f}(\mathbf{x})$, find $\mathbf{x} = \mathbf{x}(t)$.	Given $\{\mathbf{x}(t_n)\}_{n=0}^{N}$, find $\boldsymbol{f} = \boldsymbol{f}(\mathbf{x})$.
Type	Forward problem	Inverse problem
Consistency	$\rho(1) = 0, \rho'(1) = \sigma(1)$	$\rho(1) = 0, \rho'(1) = \sigma(1)$
Stability	Dalhquist root condition on ρ	New root conditions on σ (or $\hat{\sigma}$)
Example	BDF-M ($M \leq 6$), AM, AB	BDF, AM-0, AM-1, AB-M ($M \leq 6$)

when $M = 1$, as predicted in section 5. Meanwhile, also consistent with the analysis of secton 5, AB and BDF are robust for the long-time dynamics discovery—yielding a constant error for fixed mesh as T increases and a decreasing error for larger M.

7. Conclusions and future steps. In this paper, we extend the foundational work of solving ordinary differential equations using LMMs to the problem of dynamics discovery. We introduced refined notions of consistency, stability, and convergence for discovery based on classical definitions, and we showed how three prominent schemes—AB, AM, and BDF—may or may not be convergent numerical methods for dynamics discovery in general. To do so, we first derive algebraic criteria to determine the consistency and stability of the LMM, in a spirit similar to the counterpart for the classical theory. The key difference lies in the characteristic polynomial of attention; instead of the root condition for the first characteristic polynomial, as classically attributed to LMMs as time integrators, stability for discovery of dynamics is attributed to root conditions on the second characteristic polynomial. While the conditions are trivial for the BDF class, their validity in the case of AM schemes requires the study of some new properties of the Lagrange interpolants. The case of AB, at the present, has to be investigated computationally. Numerical results are presented to show agreement with the theoretical findings. In conclusion, we find theoretically and numerically that the systems for BDF-M for all $M \in \mathbb{N}$, AB for $1 \leq M \leq 6$, and AM-M for $M = 0$ and 1 are all convergent, while AB-M for $7 \leq M \leq 10$ and AM-M for $M \geq 2$ are not, as summarized in Table 2. These conclusions are drawn provided some initial data on the dynamics are supplied. Modifications need to be made, as discussed in section 4.4, if other types of additional data on the dynamics are provided. LMM schemes are well-studied for the forward problem in numerical analysis. As such tools, they can be useful to the subject of machine learning. For example, they can be applied to the design and training of neural networks that are seen as discrete forms of dynamic systems [54, 7, 49]. Different from such applications, the new study given here is motivated by recent interest in using machine learning [4, 17, 34, 35, 47] to formalize a variety of inverse problems such as learning dynamics using classical discretization techniques like LMMs. The change of the problem type from forward integration to inverse learning leads to a different mathematical theory as illustrated in Table 2.[2] Note that in particular, BDF provides a class of methods convergent for integrating and learning dynamics, while not all AB and AM methods can share the same conclusion. Our framework can be applied to check on other LMMs besides these examples. Furthermore, it will be interesting to explore if there are systematic

[2]Note that there are several different versions of consistency and stability of LMM based dynamics discovery, which also affect the order of convergence; see the discussions in Theorem 3.14.

ways to generate broader classes of LMMs good for both tasks of model-based time integration and data-driven learning.

As discussed in section 3.2, our current study assumes the best possible case that the exact states along with suitable approximations to the initial dynamics are all given, together with the assumption that the neural network representation can produce zero residual for the LMM dynamics. While this setting is highly idealized, based on the conclusions drawn, we can speculate about the impact on the properties of stability and convergence caused by different choices of time discretization schemes for a more informed attempt at discovery of unknown dynamics in more practical settings. The latter leads to many interesting issues to be considered in the future. For instance, instead of assuming only data on the state with a loss function $\mathcal{T}(\tilde{\mathbf{x}}, \tilde{\boldsymbol{f}}, \boldsymbol{f}_{NN})$, we may consider a more general loss function with data on the state and dynamics, i.e., $\mathcal{T}(\tilde{\mathbf{x}}, \tilde{\boldsymbol{f}}, \boldsymbol{f}_{NN}, \hat{\boldsymbol{f}}, \hat{\mathbf{x}})$, given by

$$\underbrace{\ell_1(\hat{\mathbf{x}}, \hat{\boldsymbol{f}})}_{\text{dynamics conformity}} + \underbrace{\ell_2(\tilde{\mathbf{x}}, \tilde{\boldsymbol{f}}, \boldsymbol{f}_{NN}, \hat{\boldsymbol{f}}) + \ell_3(\tilde{\mathbf{x}}, \hat{\mathbf{x}})}_{\text{data fidelity}} + \underbrace{\mathcal{R}_1(\hat{\mathbf{x}}) + \mathcal{R}_2(\hat{\boldsymbol{f}}) + \mathcal{R}_3(\boldsymbol{f}_{NN})}_{\text{regularization}}.$$

For LMMs with grid functions, the loss ℓ_1 associated with dynamics conformity comes from the discretization (3.1), and $\tilde{\boldsymbol{f}} \in \Gamma_h[a, b]$, the space of grid functions. The total loss can be taken as an expectation over training samples and minimized to obtain some optimal representation of the state or dynamics. LMNet is an example where the conformity term is minimized over parameterized neural networks of various types, so that $\hat{\boldsymbol{f}} \equiv \boldsymbol{f}_{NN}$, as studied in [39, 41, 52, 57].

Whenever the term involving the LMM residual is accounted for, the framework developed in this paper would be relevant. For stable LMMs considered here, one may expect that it may be possible to extend the convergence results for exact and complete data if the set of neural networks can satisfy some universal approximation properties. The convergence would be expected to be in the sense of function approximations which would imply a good generalization error, at least among suitable classes of smooth dynamic systems. For systems displaying chaotic behavior and sharp transitions, new ideas are likely needed in order to ensure accurate discovery of the underlying complex dynamics.

In this more general setting, neural network representations may also provide implicit regularization of the learned dynamics so that unstable LMMs could potentially be stabilized. However, regularization likely produces additional consistency error so the convergence has to be more carefully examined. Moreover, we may consider compressed representation and treat incomplete data by promoting sparsity or exploring the use of partial physics as regularization to achieve physics-informed and data-driven discovery of the dynamics. Finally, there are many avenues of exploration to extend the results reported here. Some interesting topics for future studies include

1. the effects of regularization by specifying various forms of the regularization terms $\mathcal{R}_1$ and $\mathcal{R}_2$, such as those promoting smoothness, sparsity, low dimensionality, and extending the above tasks for study of the dynamics discovery problem with incomplete and uncertain data,
2. different reduced-order models via choice of constrained representations on the dynamics or the state variables or both [3, 57],
3. extension of the stability framework to incorporate other multistep and multistage schemes such as predictor-corrector, Milne, and Runge–Kutta [45],
4. derivation of a general class of LMMs that are convergent for both the forward problem of time integration and the backward problem of dynamics discovery,

5. the errors in numerically *integrated* states based on learned dynamics [41],
6. distributed dynamic systems such as time-dependent PDEs and examining the additional effect due to spatial discretization,
7. generalizing to the study of dynamics for a suitable set of initial conditions.

Naturally, learning dynamics has strong connections to the subject of time-series prediction using deep learning [8, 21, 25, 51]. Our current work here may motivate further rigorous numerical analysis studies in such a direction as well. To conclude, we see from this study that there are many new challenges in physics-based and data-driven modeling and simulations warranting further numerical analysis research.

Acknowledgments. The authors would like to thank the CM3 group at Columbia University for invigorating discussions, Wen Ding for his stimulating suggestions, and the referees and Associate Editor for their valuable comments.

REFERENCES

[1] R. P. Agarwal, *Difference Equations and Inequalities: Theory, Methods, and Applications*, CRC Press, Boca Raton, FL, 2000.
[2] K. Atkinson, W. Han, and D. E. Stewart, *Numerical Solution of Ordinary Differential Equations*, Pure Appl. Math. 108, Wiley (Hoboken), Hoboken, NJ, 2011.
[3] K. Bhattacharya, B. Hosseini, N. B. Kovachki, and A. M. Stuart, *Model Reduction and Neural Networks for Parametric PDEs*, preprint, arXiv:2005.03180, 2020.
[4] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 2006.
[5] S. L. Brunton and J. N. Kutz, *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*, Cambridge University Press, Cambridge, 2019.
[6] S. L. Brunton, J. L. Proctor, and J. N. Kutz, *Discovering governing equations from data by sparse identification of nonlinear dynamical systems*, Proc. Natl. Acad. Sci. USA, 113 (2016), pp. 3932–3937.
[7] R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, *Neural ordinary differential equations*, in Advances in Neural Information Processing Systems, 2018, pp. 6571–6583.
[8] J. T. Connor, R. D. Martin, and L. E. Atlas, *Recurrent neural networks and robust time series prediction*, IEEE Trans. Neural Networks, 5 (1994), pp. 240–254.
[9] D. M. Creedon and J. J. Miller, *The stability properties OFQ-step backward difference schemes*, BIT, 15 (1975), pp. 244–249.
[10] C. W. Cryer, *On the instability of high order backward-difference multistep methods*, BIT, 12 (1972), pp. 17–25.
[11] G. Dahlquist, *Convergence and stability in the numerical integration of ordinary differential equations*, Math. Scand., 4 (1956), pp. 33–53.
[12] G. G. Dahlquist, *A special stability problem for linear multistep methods*, BIT, 3 (1963), pp. 27–43.
[13] R. C. Dorf and R. H. Bishop, *Modern Control Systems*, Pearson, London, 2011.
[14] C. Fredebeul, *A-BDF: A generalization of the backward differentiation formulae*, SIAM J. Numer. Anal., 35 (1998), pp. 1917–1938.
[15] W. Gautschi, *Numerical Analysis*, Springer, New York, 1997.
[16] H. H. Goldstine, *A History of Numerical Analysis from the* 16*th through the* 19*th Century*, Vol. 2, Springer, New York, 2012.
[17] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, Cambridge, MA, 2016.
[18] 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*, preprint, arXiv:1910.12680, 2019.
[19] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations* II: *Stiff and Differential-Algebraic Problems*, Springer, New York, 1996.
[20] J. Han, A. Jentzen, and E. Weinan, *Solving high-dimensional partial differential equations using deep learning*, Proc. Natl. Acad. Sci. USA, 115 (2018), pp. 8505–8510.
[21] M. Han, J. Xi, S. Xu, and F.-L. Yin, *Prediction of chaotic time series based on the recurrent predictor neural network*, IEEE Trans. Signal Process., 52 (2004), pp. 3409–3416.
[22] P. Henrici, *Discrete Variable Methods in Ordinary Differential Equations*, Wiley, Hoboken, NJ, 1962.

[23] M. I. Jordan and T. M. Mitchell, *Machine learning: Trends, perspectives, and prospects*, Science, 349 (2015), pp. 255–260.
[24] S. H. Kang, W. Liao, and Y. Liu, *Ident: Identifying Differential Equations with Numerical Time Evolution*, preprint, arXiv:1904.03538, 2019.
[25] F. Karim, S. Majumdar, H. Darabi, and S. Chen, *LSTM fully convolutional networks for time series classification*, IEEE Access, 6 (2017), pp. 1662–1669.
[26] I. G. Kevrekidis, C. W. Rowley, and M. O. Williams, *A kernel-based method for data-driven Koopman spectral analysis*, J. Comput. Dyn., 2 (2016), pp. 247–265.
[27] Y. Khoo, J. Lu, and L. Ying, *Solving Parametric PDE Problems with Artificial Neural Networks*, preprint, arXiv:1707.03351, 2017.
[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, *Imagenet classification with deep convolutional neural networks*, in Proceedings of Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[29] Y. LeCun, Y. Bengio, and G. Hinton, *Deep learning*, Nature, 521 (2015), pp. 436–444.
[30] Z. Long, Y. Lu, and B. Dong, *PDE-Net 2.0: Learning PDEs from data with a numeric-symbolic hybrid deep network*, J. Comput. Phys., 399 (2019), 108925.
[31] F. Lu, M. Zhong, S. Tang, and M. Maggioni, *Nonparametric inference of interaction laws in systems of agents from trajectory data*, Proc. Natl. Acad. Sci. USA, 116 (2019), pp. 14424–14433.
[32] C. Ma, J. Wang, and W. E, *Model Reduction with Memory and the Machine Learning of Dynamical Systems*, arXiv:1808.04258, 2018.
[33] D. Mayers and E. Süli, *An Introduction to Numerical Analysis*, Cambridge University Press, Cambridge, 2003.
[34] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*, MIT Press, Cambridge, MA, 2018.
[35] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*, MIT Press, Cambridge, MA, 2012.
[36] F. Odeh and W. Liniger, *A note on unconditional fixed-h stability of linear multistep formulae*, Computing, 7 (1971), pp. 240–253.
[37] S. Pan and K. Duraisamy, *Data-driven discovery of closure models*, SIAM J. Appl. Dyn. Syst., 17 (2018), pp. 2381–2413.
[38] E. Qian, B. Kramer, B. Peherstorfer, and K. Willcox, *Lift & learn: Physics-informed machine learning for large-scale nonlinear dynamical systems*, Phys. D, 406 (2020), 132401.
[39] T. Qin, K. Wu, and D. Xiu, *Data driven governing equations approximation using deep neural networks*, J. Comput. Phys., 395 (2019), pp. 620–635.
[40] M. Raissi, *Deep hidden physics models: Deep learning of nonlinear partial differential equations*, J. Mach. Learn. Res., 19 (2018), pp. 932–955.
[41] M. Raissi, P. Perdikaris, and G. E. Karniadakis, *Multistep Neural Networks for Data-driven Discovery of Nonlinear Dynamical Systems*, preprint, arXiv:1801.01236, 2018.
[42] M. Raissi, P. Perdikaris, and G. E. Karniadakis, *Numerical Gaussian processes for time-dependent and nonlinear partial differential equations*, SIAM J. Sci. Comput., 40 (2018), pp. A172–A198, https://doi.org/10.1137/17M1120762.
[43] M. Raissi, P. Perdikaris, and G. E. Karniadakis, *Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations*, J. Comput. Phys., 378 (2019), pp. 686–707.
[44] S. H. Rudy, S. L. Brunton, J. L. Proctor, and J. N. Kutz, *Data-driven discovery of partial differential equations*, Sci. Adv., 3 (2017), e1602614.
[45] S. H. Rudy, J. N. Kutz, and S. L. Brunton, *Deep learning of dynamics and signal-noise decomposition with time-stepping constraints*, J. Comput. Phys., 396 (2019), pp. 483–506.
[46] M. Schmidt and H. Lipson, *Distilling free-form natural laws from experimental data*, Science, 324 (2009), pp. 81–85.
[47] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press, Cambridge, 2014.
[48] S. Strelitz, *On the Routh-Hurwitz problem*, Amer. Math. Monthly, 84 (1977), pp. 542–544.
[49] Q. Sun, Y. Tao, and Q. Du, *Stochastic Training of Residual Networks: A Differential Equation Viewpoint*, preprint, arXiv:1812.00174, 2018.
[50] Y. Sun, L. Zhang, and H. Schaeffer, *NEUPDE: Neural Network Based Ordinary and Partial Differential Equations for Modeling Time-Dependent Data*, preprint, arXiv:1908.03190, 2019.
[51] Y. Tao, L. Ma, W. Zhang, J. Liu, W. Liu, and Q. Du, *Hierarchical Attention-Based Recurrent Highway Networks for Time Series Prediction*, preprint, arXiv:1806.00685, 2018.

[52] 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*, 2019, https://arxiv.org/abs/1904.04058.
[53] M. Wang, H.-X. Li, X. Chen, and Y. Chen, *Deep learning-based model reduction for distributed parameter systems*, IEEE Trans. Syst. Man Cybern. Syst., 46 (2016), pp. 1664–1674.
[54] E. Weinan, *A proposal on machine learning via dynamical systems*, Commun. Math. Stat., 5 (2017), pp. 1–11.
[55] M. O. Williams, I. G. Kevrekidis, and C. W. Rowley, *A data–driven approximation of the Koopman operator: Extending dynamic mode decomposition*, J. Nonlinear Sci., 25 (2015), pp. 1307–1346.
[56] K. Wu and D. Xiu, *Data-Driven Deep Learning of Partial Differential Equations in Modal Space*, preprint, arXiv:1910.06948, 2019.
[57] X. Xie, G. Zhang, and C. G. Webster, *Non-Intrusive Inference Reduced Order Modeling of Fluid Dynamics Using Linear Multistep Network*, preprint, arXiv:1809.07820, 2018.
[58] Y. Zhu and N. Zabaras, *Bayesian deep convolutional encoder–decoder networks for surrogate modeling and uncertainty quantification*, J. Comput. Phys., 366 (2018), pp. 415–447.