Assimilation-based Learning of Chaotic Dynamical Systems from Noisy and Partial Data

Duong Nguyen, Said Ouala, Lucas Drumetz, Ronan Fablet

ASSIMILATION-BASED LEARNING OF CHAOTIC DYNAMICAL SYSTEMS FROM NOISY AND PARTIAL DATA

Duong Nguyen, Said Ouala, Lucas Drumetz, Ronan Fablet

IMT Atlantique, Lab-STICC, F-29238 Brest, France

ABSTRACT

Despite some promising results under ideal conditions (i.e. noise-free and complete observation), learning chaotic dynamical systems from real life data is still a very challenging task. We propose a novel framework, which combines data assimilation schemes and neural network representation, namely Auto-Encoders and Ensemble Kalman Smoother, to learn the governing equations of dynamical systems. By treating the learning as a Bayesian estimation problem, our framework can deal with noisy and partial observations. Experiments on the chaotic Lorenz–63 dynamics with different noise settings demonstrate the advantages of our method over the state-of-the-art.

***Index Terms*—** dynamical systems, data assimilation, neural networks.

1. INTRODUCTION

Dynamical systems are at the core of numerous scientific fields, among which we may cite geosciences, aerodynamics, fluid dynamics, etc. Classically, the determination of the governing laws of a given system, usually stated as Ordinary Dynamical Equations (ODE) or Partial Differential Equations (PDE), combines mathematical derivation based on physical principles and some experimental validations [1, 2, 3, 4]. An alternative approach is to learn the governing equations from their output, i.e. the observations [5, 6]. Recently, such data-driven paradigms have experienced a rebirth thanks to advances in modern machine learning [7] and the availability of more and more data. Under ideal conditions, with high signal-to-noise ratio (SNR) and high sampling frequency, numerous methods have successfully captured the hidden dynamics of systems [8, 9, 10, 11, 12, 13]. However, real life data usually come with noise and can be irregularly sampled, both in spatial and temporal sense, makes learning dynamical systems an extremely difficult task, and all the methods above fail.

In this paper, we show that by adding an inference scheme to any existing model, we can significantly improve its learning capacity. This framework combines the advances of state-of-the-art machine learning—neural networks—and data assimilation schemes for the problem of learning dynamical systems. The identification of dynamical models is embedded into a Bayesian analysis problem, and is explicitly decomposed into two parts: an assimilation part, where an inference scheme retrieves the true hidden states from noisy observations and a learning part where a neural network learns the dynamics from the states identified by the assimilation step. Our experiments support an extension of learning capacities for significant noise levels and irregularly-sampled data, which are widely-encountered for real-world dynamical datasets.

The paper is organized as follows. In Section 2, we formulate the problem of learning non-linear dynamical systems. State-of-the-arts are briefly reviewed in Section 3. Section 4 presents the details of proposed framework, followed by the experiments and results in Section 5. Finally, we end the paper with conclusions and perspectives for future work in Section 6.

2. PROBLEM FORMULATION

The identification of data-driven representations of dynamical systems amounts to determining a data-driven computational representation which approximates the true dynamical model typically stated as an ODE (Ordinary Differential Equation):

$$\frac{d\mathbf{x}_t}{dt} = \mathcal{F}\big(\mathbf{x}_t\big) + \eta_t \tag{1}$$

where $\mathcal{F}$ is the unknown dynamical model of a multi-dimensional state $\mathbf{x}_t$ and η_t is a zero-mean additive noise accounting for neglected physics and/or numerical approximations. State-of-the-art data-driven schemes typically exploit observation data $\mathbf{y}_{t_i}$ as:

$$\mathbf{y}_{t_i} = \Phi_{t_i}\big(\mathcal{H}\big(\mathbf{x}_{t_i}\big) + \epsilon_t\big) \tag{2}$$

where $\mathcal{H}$ is the observation operator, usually known, ϵ_t is a zero-mean noise process and $\{t_i\}_i$ refers to the time sampling, typically a regular high-frequency sampling such that $t_i = t_0 + i.\delta$ with respect to a time resolution δ and a reference starting point t_0. We introduce a masking operator Φ_{t_i} to account for the fact that observation $\mathbf{y}_{t_i}$ may not be available at all time t_i ($\Phi_{t_i,j} = 0$ if the j^{th} variable of $\mathbf{y}_{t_i}$ is missing).

This work was supported by Labex Cominlabs (grant SEACS), CNES (grant OSTST-MANATEE), Microsoft (AI EU Ocean awards) and by MESR, FEDER, Région Bretagne, Conseil Général du Finistère, Brest Métropole and Institut Mines Télécom in the framework of the VIGISAT program managed by "Groupement Bretagne Télédétection" (BreTel).

3. RELATED WORK

Learning dynamical systems is a topic that has been studied for decades. Before the era of deep learning, most schemes used Expectation-Maximization (EM)-like iterative learning procedures [6, 14, 15]. These models use a Kalman filter-based data assimilation schemes in the E-step to estimate the true hidden states $\mathbf{x}_t$, then perform an analytical calculation to retrieve model parameters in the M-step. However, such approaches apply only on simple distributions and processes whose analytic form is known. Therefore, their applicability may be fairly limited.

Recently, deep learning [16] has been leveraged across many domains, including model identification. [12] used DenseNet, [13] used ResNet, [17] used LSTM to identify the nonlinear dynamical systems by minimizing the short-term prediction error. These methods exploit the power of neural networks to overcome the difficulties of modeling non-linearities. Although neural networks are very powerful and can theoretically approximate any function [18], no regularization techniques have been proved effective to learn dynamical systems. Problems arise when the observations are partial, irregular or in case of a high level of noise. When the observations are highly corrupted by noise, using the short-term prediction error as the objective function would very likely lead to overfitting the data. Besides, neural networks, in general, do not have an efficient way to deal with irregularly-sampled data.

There are two special methods that do not use neural networks: the Analog Forecasting (AF) [8] and the sparse regression (Sparse Identification of Nonlinear Dynamics-SINDy) [9]. Analog forecasting is a non-parametric model that "learns by heart" the dynamics from a reference catalog. It predicts the evolution by looking for the most similar states in the catalog and combines their successors into the predicted state. Since AF is a k-Nearest Neighbors based method, it does not work well in high-dimensional spaces. SINDy aims to identify the analytic form of the dynamics by performing a sparse regression on a basis formed by candidate functions. This method works extremely well when the observations are complete and clean. However, an accurate estimate of the gradient of the data is crucial. When the observation is highly noisy, partial or irregular, gradients may be poorly estimated and SINDy may likely fail.

4. PROPOSED FRAMEWORK

In this Section, we present the proposed framework for the data-driven identification of neural-network representations of dynamical systems from noisy and irregularly-sampled observations. Rather than considering the learning as a short-term prediction error optimization problem, the proposed framework is inspired by the Bayesian formulation used for data assimilation. It provides neural networks the ability to deal with noise and irregularities.

Formally, we consider a discrete state-space formulation, which amounts to reformulating Eqs. (1) and (2) as follows:

$$\mathbf{x}_{t+\delta} = f_\delta\big(\mathbf{x}_t\big) + \omega_{t+\delta} \tag{3}$$

$$\mathbf{y}_t = \Phi_t\big(\mathcal{H}\big(\mathbf{x}_t\big) + \epsilon_t\big) \tag{4}$$

Where $\mathbf{x}_{t+\delta}$ results from an integration of operator $\mathcal{F}$ from state $\mathbf{x}_t$: $f_\delta\big(\mathbf{x}_t\big) = \mathbf{x}_t + \int_t^{t+\delta} \mathcal{F}\big(\mathbf{x}_t\big)dt$. ω_t and ϵ_t represent the uncertainty of the model and the observation, respectively. For the sake of simplicity, δ is arbitrarily set to 1 without any loss of generality. Within this Bayesian setting, Eqs. (3) and (4) relate respectively to the dynamical prior $p\big(\mathbf{x}_{t+\delta}|\mathbf{x}_t\big)$ and the observation likelihood $p\big(\mathbf{y}_t|\mathbf{x}_t\big)$.

Using a fixed integration scheme, the problem results in maximizing $\ln p_\theta(\mathbf{y}_{1:T}) = \ln \int p_\theta(\mathbf{y}_{1:T}, \mathbf{x}_{1:T}) d\mathbf{x}_{1:T}$ for a sequence of T observations $\mathbf{y}_{1:T} = \mathbf{y}_1..\mathbf{y}_T$. Apart from simple cases, $\ln p_\theta(\mathbf{y}_{1:T})$ is intractable [19]. We usually maximize the following Evidence Lower BOund (ELBO) instead:

$$\mathcal{L}(\mathbf{y}_{1:T}, p_\theta, q) = \int q(\mathbf{x}_{1:T}|\mathbf{y}_{1:T}) \ln \frac{p_\theta(\mathbf{y}_{1:T}, \mathbf{x}_{1:T})}{q(\mathbf{x}_{1:T}|\mathbf{y}_{1:T})} d\mathbf{x}_{1:T} \tag{5}$$

with arbitrary function q_ϕ. Here we focus on the family of distributions that can be factored over t:

$$p_\theta(\mathbf{y}_{1:T}, \mathbf{x}_{1:T}) = p_\theta(\mathbf{x}_1) \prod_{t=1}^{T-1} p_\theta(\mathbf{x}_{t+1}|\mathbf{x}_t) \prod_{k=1}^{T} p_\theta(\mathbf{y}_k|\mathbf{x}_k) \tag{6}$$

$$q(\mathbf{x}_{1:T}|\mathbf{y}_{1:T}) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{y}_{1:T}) \tag{7}$$

In geosciences, aerodynamics and fluid dynamics, dynamical systems are usually modeled as deterministic functions [2, 4]. If we restrict q to Dirac distributions (in other words, we use a Maximum A Posteriori—MAP estimator), Eq. (5) becomes:

$$\mathcal{L}(\mathbf{y}_{1:T}, p_\theta, q) = \ln p_\theta(\mathbf{x}_1^*) + \sum_{t=1}^{T-1} \ln p_\theta(\mathbf{x}_{t+1}^*|\mathbf{x}_t^*) + \sum_{k=1}^{T} \ln p_\theta(\mathbf{y}_k|\mathbf{x}_k^*) \tag{8}$$

with $\mathbf{x}_t^* = \mathbb{E}\left[q(\mathbf{x}_t|\mathbf{y}_{1:T})\right]$.

Assuming that the covariance matrices of the observation operator are constant and diagonal, we define a new loss function:

$$\mathcal{L}'(\mathbf{y}_{1:T}, p_\theta, q) = \sum_{t=1}^{T-1} ||f(\mathbf{x}_t^*) - \mathbf{x}_{t+1}^*||_2^2 + \lambda \sum_{t=1}^{T} ||\Phi_t(\mathcal{H}(\mathbf{x}_t^*)) - \mathbf{y}_t||_2^2 \tag{9}$$

where $||.||_2$ is the Euclidean distance. This formulation comprises a short-term prediction error $||f(\mathbf{x}_t^*) - \mathbf{x}_{t+1}^*||_2^2$, which is the loss function widely used in most of data-driven dynamical systems identification models [10, 11, 12, 13]. We may stress that here the prediction is applied on the true states $\mathbf{x}_t$ instead of on the observations $\mathbf{y}_t$. As such, the model does not overfit the noise in data. The second term on the right hand side of Eq. (9) is analogous to the innovation (the measurement of pre-fit residual) in data assimilation schemes [20]. It is trivial to prove that minimizing the loss function in Eqs. (9) amounts to maximizing a lower bound in Eq. (5) with the two assumptions above, although this bound is infinitely loose. However, on one hand, tighter variational bounds are not necessarily better [21]. And, on the other hand, this simplification significantly reduces the computational cost.

So far we have presented the mathematical assumptions and derived the formulations that support the combination of a short-term predicting model for learning dynamical systems and a data assimilation scheme. We may sum up the the key idea as follows. A data-driven method for dynamical system identification should involve two key components: (i) a data assimilation stage that retrieves from the observations the true hidden states where the dynamics lie on, (ii) a learning stage that learns the dynamics of the true states. This framework is general and can be applied for many kinds of models. f_δ can be modeled by any current state-of-the-art dynamical systems identification model. We here focus on neural-network-based models, because of their generality and computational efficiency. For the assimilation scheme, we investigate two strategies: the Ensemble Kalman Smoother (EnKS) and LSTM Auto-Encoder (LSTMAE), the former being among the state-of-the-art schemes in data assimilation and the latter among the state-of-the-art inference schemes in machine learning.

5. EXPERIMENTS AND RESULTS

The proposed methodology was tested on the Lorenz–63 dynamics [1]. They are representative of chaotic dynamics, which make them appealing for our benchmarking experiments. We examined the learning performance under significant noise level with partial and irregular sampling of the observations. We consider benchmarking experiments with state-of-the-art methods, namely Analog forecasting (AF) [8], Sparse regression model (SINDy) [9] and Bilinear residual Neural Network (BiNN) [11].

5.1. Set up

We simulated a Lorenz-63 state sequence of length 4000 using the LOSDA (Livermore Solver for Ordinary Differential Equations) ODE solver [22] with an integration step of 0.01, then added Gaussian noise with several variance levels σ^2. We evaluated the considered schemes given the noisy and sub-sampled (possibly irregularly) training data. This means, $\mathcal{H}$ was an identity operator, ϵ_t was a zero-mean Gaussian white noise.

We applied AF, SINDy and BiNN using the setting proposed in the related original papers. As stated above, our methodology can be applied to any neural-network-based model. Here, we chose the BiNN to model $\mathcal{F}$, as this architecture embeds the true parametrization of the Lorenz-63 system. The integration scheme was a neural network implementation of the Runge-Kutta 4 as in [11]. For the EnKS, we chose an ensemble of 50 members. For the LSTMAE, the approximate posterior $q(\mathbf{x}_{1:T}|\mathbf{y}_{1:T})$ was modeled by a 2-layer bi-directional LSTM with a size of 9. Both the encoder and the decoder were modeled by a fully connected network, with one hidden layer of size 7. We used RMSprop as the optimizer, with a learning rate of $1e-3$. λ was set to 0.01.

5.2. Identification with noisy observations

We first assess the identification performance with noisy observations only, which means that the masking operator Φ_t is an identity matrix at all time steps. We evaluated both the short-term prediction and the capacity of maintaining the very-long-term topology through the first Lyapunov exponent λ_1 calculated in a forecasting sequence of length of 10000 time steps (the true λ_1 of the Lorenz-63 system is 0.91 [3]).

As shown in Table 1, both the implementations of the proposed framework (BiNN with EnKS, denoted as BiNN-EnKS and BiNN with LSTMAE, denoted as BiNN-LSTMAE) outperform existing methods. BiNN-EnKS gives the best forecasting when the noise level is small, however, when the noise level increases, the forecasting gradually becomes worse. This is because the increase of noise level leads to the increase of uncertainty (error covariance). On the other hand, BiNN-LSTMAE can maintain its performance level even when the noise level is very high. We believe that this relates to the ability of LSTM-based models to capture long-term patterns in the data. The Lorenz-63 system has a positive Lyapunov exponent, meaning that any small initial difference model grows exponentially in the long-term. However, Lorenz-63 sequences clearly involve a pattern, referred to an attractor. As discussed in [9] and [10], for dynamical models identification, the most important criterion is the ability to maintain this topology in very-long forecasting sequences. Fig. 1 depicts the attractor of the sequences generated by the models in Table 1. Topology-wise, we can see that the attractors generated by BiNN-EnKS and the BiNN-LSTMAE are visually better.

When the training data are clean, many data-driven methods can successfully identify the underlying dynamics. For example, BiNN and BiNN-LSTMAE have similar result when $\sigma^2 = 0.5$. However, when the data are very noisy, AF, SR and BiNN schemes fail. The explicit assimilation schemes help disentangle the true latent states from noisy observations. As a result, the learning part is weakly affected by the noise level compared with the learning under ideal conditions.

Table 1: Short-term forecasting error and very-long-term forecasting topology of data-driven models learned on noisy Lorenz-63 data.

Model		σ^2 0.5	2	8	32
AF	t_0+4	0.245	0.698	2.213	3.887
	λ_1	-1.356	-2.496	-1.900	32.432
SiNDy	t_0+4	0.037	0.104	0.326	0.933
	λ_1	0.890	0.876	-0.367	nan
BiNN	t_0+4	0.043	0.061	0.296	0.773
	λ_1	0.912	0.833	nan	-0.014
BiNN-EnKS	t_0+4	**0.013**	**0.027**	0.055	0.156
	λ_1	0.859	0.842	0.878	0.892
BiNN-LSTMAE	t_0+4	**0.013**	0.028	**0.030**	**0.047**
	λ_1	0.899	0.872	0.919	0.912

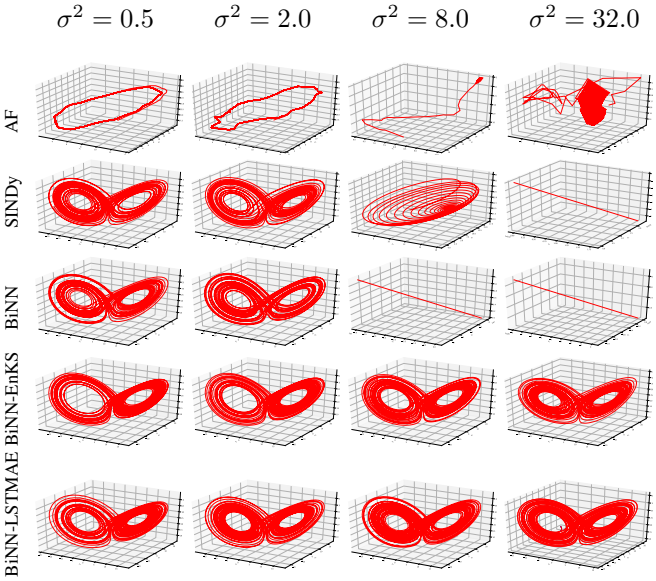

Fig. 1: Attactors generated by models trained on noisy data.

5.3. Identification with noisy and partial observations

In many application domains of dynamical systems, one cannot monitor data at high frequency continuously. This section evaluates the performance of the proposed framework on not only noisy but also partial observations. Specifically, 80% of data were missing. We considered two scenarios for partial observations: i) s1: one complete observation is observed every 8 time steps, regularly; and ii) s2: the data are observed irregularly, along multivariate and time dimensions, but overall, 20% of data are observed.

We used the same evaluation criteria as those in Section 5.2. The results are shown in Table. 2 and Fig. 2. Unless the noise level is extremely high, both BiNN_EnKS and BiNN_LSTMAE are able to capture the dynamical laws the govern the observations.

Table 2: Short-term forecasting error and very-long-term forecasting topology of data-driven models learned on noisy and partial Lorenz-63 data.

Model		σ^2 0.5	2	8	32
BiNN-EnKS_s1	t_0+4	0.015	0.038	0.066	0.186
	λ_1	0.903	0.896	0.744	0.894
BiNN-LSTMAE_s1	t_0+4	0.057	0.060	0.165	0.232
	λ_1	0.899	0.911	0.923	0.468
BiNN-EnKS_s2	t_0+4	0.065	0.075	0.156	0.149
	λ_1	0.894	0.758	0.475	0.658
BiNN-LSTMAE_s2	t_0+4	0.036	0.117	0.253	0.318
	λ_1	0.908	0.868	0.854	0.312

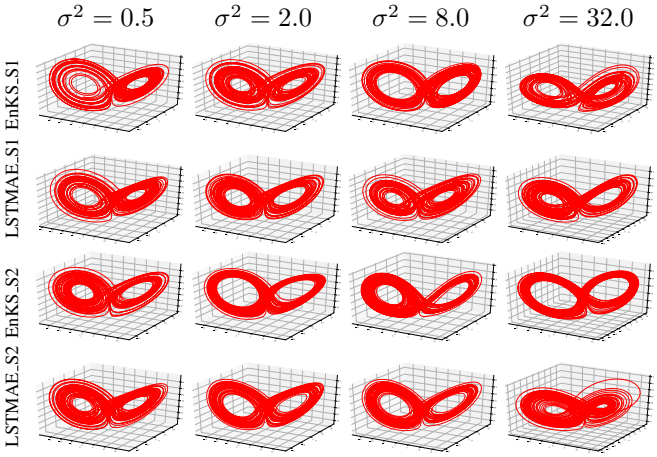

Fig. 2: Attractors generated by models trained on partially observed data scenario 1 and 2. (The terms "BiNN" in the name of the models are omitted)

6. CONCLUSIONS

This paper has showed that by explicitly adding an assimilation scheme, we could significantly improve the capacity of neural networks to learn chaotic dynamical systems. Our proposed framework combines state-of-the-art data assimilation schemes and recent advances in neural networks for the data-driven identification of governing equations of non-linear dynamical systems. Different communities may value our contributions for different aspects. For the data assimilation community, we introduce neural networks as a means to go beyond the limit of using analytic functions/processes, such as the nonlinearity. For deep learning practitioners, our experiments point out that assimilation schemes can be loosely considered as a regularization technique to prevent overfitting.

A number of open problems remain for future work. We use the two hypotheses to derive the short-term prediction error term used in existing dynamics learning methods. However, the first one (q is a Dirac distribution) can be relaxed to model stochastic systems, the second one (constant diagonal variance) can be relaxed to estimate the observation errors.

7. REFERENCES

[1] E. N. Lorenz, "Deterministic Nonperiodic Flow," *Journal of the Atmospheric Sciences*, vol. 20, no. 2, pp. 130–141, Mar. 1963. [Online]. Available: https://journals.ametsoc.org/doi/abs/10.1175/1520-0469(1963)020%3C0130:dnf%3E2.0.CO;2

[2] R. C. Hilborn, *Chaos and nonlinear dynamics: an introduction for scientists and engineers*. Oxford University Press on Demand, 2000.

[3] J. C. Sprott and J. C. Sprott, *Chaos and time-series analysis*. Citeseer, 2003, vol. 69.

[4] M. W. Hirsch, S. Smale, and R. L. Devaney, *Differential equations, dynamical systems, and an introduction to chaos*. Academic press, 2012.

[5] Z. Ghahramani and G. E. Hinton, "Variational learning for switching state-space models," *Neural Computation*, vol. 12, pp. 963–996, 1998.

[6] ——, "Parameter estimation for linear dynamical systems," Technical Report CRG-TR-96-2, University of Totronto, Dept. of Computer Science, Tech. Rep., 1996.

[7] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, Jul. 2015. [Online]. Available: http://science.sciencemag.org/content/349/6245/255

[8] R. Lguensat, P. Tandeo, P. Ailliot, M. Pulido, and R. Fablet, "The Analog Data Assimilation," *Monthly Weather Review*, vol. 145, no. 10, pp. 4093 – 4107, Oct. 2017. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01609141

[9] 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*, vol. 113, no. 15, pp. 3932–3937, Apr. 2016. [Online]. Available: https://www.pnas.org/content/113/15/3932

[10] J. Pathak, Z. Lu, B. R. Hunt, M. Girvan, and E. Ott, "Using Machine Learning to Replicate Chaotic Attractors and Calculate Lyapunov Exponents from Data," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 27, no. 12, p. 121102, Dec. 2017, arXiv: 1710.07313. [Online]. Available: http://arxiv.org/abs/1710.07313

[11] R. Fablet, S. Ouala, and C. Herzet, "Bilinear residual Neural Network for the identification and forecasting of dynamical systems," *arXiv:1712.07003 [physics]*, Dec. 2017, arXiv: 1712.07003. [Online]. Available: http://arxiv.org/abs/1712.07003

[12] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Multistep Neural Networks for Data-driven Discovery of Nonlinear Dynamical Systems," *arXiv:1801.01236 [nlin, physics:physics, stat]*, Jan. 2018, arXiv: 1801.01236. [Online]. Available: http://arxiv.org/abs/1801.01236

[13] T. Qin, K. Wu, and D. Xiu, "Data Driven Governing Equations Approximation Using Deep Neural Networks," *arXiv:1811.05537 [cs, math, stat]*, Nov. 2018, arXiv: 1811.05537. [Online]. Available: http://arxiv.org/abs/1811.05537

[14] Z. Ghahramani and S. T. Roweis, "Learning Nonlinear Dynamical Systems Using an EM Algorithm," in *Advances in Neural Information Processing Systems 11*, M. J. Kearns, S. A. Solla, and D. A. Cohn, Eds. MIT Press, 1999, pp. 431–437. [Online]. Available: http://papers.nips.cc/paper/1594-learning-nonlinear-dynamical-systems-using-an-em-algorithm.pdf

[15] H. U. Voss, J. Timmer, and J. Kurths, "Nonlinear dynamical system identification from uncertain and indirect measurements," *International Journal of Bifurcation and Chaos*, vol. 14, no. 06, pp. 1905–1933, Jun. 2004. [Online]. Available: https://www.worldscientific.com/doi/abs/10.1142/S0218127404010345

[16] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[17] K. Yeo and I. Melnyk, "Deep learning algorithm for data-driven simulation of noisy dynamical system," *Journal of Computational Physics*, vol. 376, pp. 1212–1231, Jan. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0021999118306910

[18] B. C. Csji, "Approximation with artificial neural networks," *Faculty of Sciences, Etvs Lornd University, Hungary*, vol. 24, p. 48, 2001.

[19] C. M. Bishop, *Pattern recognition and machine learning*, ser. Information science and statistics. New York, NY: Springer, 2006.

[20] G. Evensen, *Data Assimilation: The Ensemble Kalman Filter*. Springer Science & Business Media, Aug. 2009, google-Books-ID: 2_zaTb_O1AkC.

[21] T. Rainforth, A. R. Kosiorek, T. A. Le, C. J. Maddison, M. Igl, F. Wood, and Y. W. Teh, "Tighter Variational Bounds are Not Necessarily Better," *arXiv:1802.04537 [cs, stat]*, Feb. 2018, arXiv: 1802.04537. [Online]. Available: http://arxiv.org/abs/1802.04537

[22] A. C. Hindmarsh, "ODEPACK, a systematized collection of ODE solvers," *IMACS Transactions on Scientific Computation*, vol. 1, pp. 55–64, 1983.