Dimensional Reduction of Dynamical Systems by Machine Learning: Automatic Generation of a Macroscopic Model

Tomoaki Nogawa

Faculty of Medicine, Toho University, 5-21-16,
Omori-Nishi, Ota-ku, Tokyo, 143-8540, Japan
(Dated: June 9, 2020)

Abstract

We propose the framework to generate a phenomenological model that extract the essence of a dynamical system with large degrees of freedom by using machine learning. For a given microscopic dynamical system, we simultaneously seek for the suitable projection to a macroscopic variable, which is supposed to be extensive, and the time proceeding equation that governs them. The utility of this method is demonstrated by the application to the elementary cellular automata.
In the study of the dynamics of a complex system with large degrees of freedom, we often use a phenomenological model with small degrees of freedom, e.g., the Lorentz model for atmospheric variability \[1\]. Once such a model is supposed, we have various analysis techniques developed in the field of so-called nonlinear dynamics. Usually, such a reduced model is empirically introduced with a drastic approximation owing to the intuition and insight of researchers. Although it is an orthodox task for statistical physicists to derive such a model ab initio from a basic microscopic model, it is impossible in most cases. The aim of this paper is to propose a generic framework to generate a macroscopic dynamical system (DS) for a given microscopic DS with help from machine learning. One of the key feature of the modern machine learning, represented by deep neural networks \[2\], is automatic extraction of feature amount of the object. In the similar manner, we try to find the suitable variable to describe the dynamics of a many-body system. The DS that the obtained variables obey is also a product of the learning.

Let us start with a microscopic DS

\[
x_{t+1} = f(x_t) \in \mathbb{R}^N, \quad N \gg 1.
\]  

(1)

The goal is obtaining a macroscopic DS in the form

\[
X(x_{t+1}) = F(X(x_t)) \in \mathbb{R}^n, \quad n \ll N,
\]  

(2)

where \(X : \mathbb{R}^N \to \mathbb{R}^n\) is a projection to macroscopic variable and \(F : \mathbb{R}^n \to \mathbb{R}^n\) is time proceeding operation. It is desired that \(X\) is interpretable for human being. For this goal, we use machine learning in which a data point is \((x_t, x_{t+1}) \in \mathbb{R}^{N \times 2}\) satisfying Eq. (1). More concretely, we seek for \(F\) and \(X\) that minimize the error functional

\[
E[F, X] := \frac{\|F(X(x_t)) - X(x_{t+1})\|^2}{n}.
\]  

(3)

Hereafter, the overline denotes the average over samples. If \(E[F, X]\) equals zero, the obtained DS is exactly closed. Otherwise, it is an approximated formula, whose precision is evaluated by the residual error. The error generally decreases with \(n\). This \(n\)-dependence would be a clue to quantify the complexity or degree of nonequilibrium of the system.

If \(X\) is given, the task is simple regression of \(F\) in the from of supervised learning. There are a lot of studies on this kind of regression of DS \[3, 4\]. If \(X\) is not given, it is not a popular problem. This is, however, similar to the finite size scaling \[5, 6\], where we seek for both how to scale the variables and what equation the scaled variables satisfy.
In the following, we employ the elementary cellular automata (ECA) [7] as an example. The system has binary variables: \( x_{it} \in \{0, 1\} \) where \( i \in \{1, \cdots , N\} \) and \( t \in \{0, 1, \cdots \} \) denote the position in one-dimensional chain and time, respectively. Time evolution is driven by the rule in which \( x_{i,t+1} \) is determined by three variables: \( (x_{i-1,t}, x_{it}, x_{i+1,t}) \). We impose the periodic boundary condition for \( i \). The number of the possible rules is \( 2^{2^3} = 256 \).

There exists indefiniteness in \( X \); If \((X, F)\) satisfies Eq. (2), \((G \circ X, G \circ F \circ G^{-1})\) with arbitrary \( G : \mathbb{R}^n \to \mathbb{R}^n \) that has an inverse function does too. To avoid this indefiniteness, we need to impose some restrictions on \( X \). Here, we suppose that each component of \( X \) takes the form

\[
X_m(x) = \frac{1}{N_b} \sum_{j=1}^{N_b} \xi_m(y_j), \quad y_j \in \mathbb{R}^b, \quad y_j^\ell = x_{b(j-1)+\ell}.
\]

(4)

Here, the system is divided into blocks with length \( b \) and \( X_m \) is given by the summation of a local \( b \)-body function \( \xi_m \). This makes \( X_m \) an extensive variable (normalized by \( N_b := \lfloor N/b \rfloor \)). In addition, we impose the normalization condition that \( X_m(x_t) = 0 \) and \( X_m(x_t)^2 = 1 \) for all \( m \). This excludes the meaningless solution of \( E[X, F] = 0: X = F = 0 \).

In the case of a binary variable, the state of each block is represented by an integer \( z_j := 1 + \sum_{\ell=1}^{b} y_{j\ell} 2^{\ell-1} \in \{1, \cdots , K\} \), where \( K := 2^b \). Thus, the local function is expressed as \( \xi_m(y_j) = \sum_{k=1}^{K} w_{mk} \delta_{z_{jk}} \), which leads to

\[
X_m(x) = \sum_{k=1}^{K} w_{mk} \bar{X}_k(x), \quad \bar{X}_k(x) := \frac{1}{N_b} \sum_{j=1}^{N_b} \delta_{z_{jk}}.
\]

\[
\iff \quad X(x) = w \bar{X}(x), \quad w \in \mathbb{R}^{n \times K}, \quad \bar{X}(x) \in \mathbb{R}^K.
\]

(5)

We can use \((\bar{X}(x_t), \bar{X}(x_{t+1}))\) as a data point instead of \((x_t, x_{t+1})\). This reduces the dimension of the data point from \( N \times 2 \) to \( K \times 2 \). In general, the residual error decreases with both \( n \) and \( b \) because these increase the degree of expression of the macroscopic DS. (Note that if \( b' \) is a divisor of \( b \), the \( b \)-body function includes the \( b' \)-body one.) There are, however, trade-off between low error and ease to understand of the meaning of the macroscopic variables for human being. Since we have at most \( K - 2 \) linear independent variables, it is waste to have \( n > K - 2 \) (Here, minus-two comes from the constraint \( \sum_{k=1}^{K} \bar{X}_k(x) = 1 \) and the normalization of \( \bar{X}_k \)). Therefore, we may not be able to obtain exact DS even for \( n \to N \) unless \( b = N \).

On the function form of \( F \), we simply suppose a \( n \)-variable polynomial up to the \( p \)-th
FIG. 1. Samples of the spatio-temporal patterns of the ECA with the rules: 40 (a), 108 (b), 110 (c) and 90 (d) for $N = 128$ and $p_0 = 0.5$. Black squares represent $x_{it} = 1$ and the white ones do $x_{it} = 1$. The horizontal and vertical directions indicate $i:1-32$ and $t:0-31$, respectively.

order as

$$F(X) = W\tilde{F}(X), \quad W \in \mathbb{R}^{n \times P}, \quad P = \frac{(p+n)!}{p!n!},$$

$$\tilde{F}_t(X) = \prod_{m=1}^{n} X_m^{a_{tm}}, \quad a_{tm} \in \{0, 1, \cdots, p\}, \quad \sum_{m=1}^{n} a_{tm} \leq p. \quad (6)$$

Eventually, the residual error is expressed as a function of the weight $(w, W) \in \mathbb{R}^{n \times (K+P)}$. In the learning process, we try to reduce the error by changing $w$ and $W$ alternately. When $w$ is fixed, the problem is linear regression, and $W$ is optimized by using the covariance matrix of $\tilde{F}$. When $W$ is fixed, we use the stochastic gradient descent method on $w$. The detail of the implementation is explained at the end of this paper.

The optimal weight depends on how data $(x_t, x_{t+1})$ are sampled. In this work, we generate a data point by starting from a random initial state $x_0$ and proceeding time up to uniform random time: $t \in \{1, \cdots, t_{\text{max}}\}$, where $t_{\text{max}}$ is a fixed parameter. The binary variable $x_{i0}$ for each $i$ is set to be 0 with probability $p_0$ and to be 1 otherwise. Here, $p_0$ is given as a uniform random number in $[0, 1)$. This nonlocal randomness is introduced in order to avoid the following situation. If $p_0$ is a constant, $\bar{X}$ is self-averaging at each $t$; $X(x_t)$ is always a univalued function of $t$ for $N \to \infty$. Thus, closed DS is obtained for arbitrary $X(x_t)$ that is a monotonic function of $t$.

In the following, we show the results for the four ECA rules (40, 108, 110 and 90) that
show distinct types of dynamics according to the classification by Wolfram [7]. Typical pattern of the time evolution is shown in Fig. 1. In this part, we note $X_m(x_t) = X_{mt}$. If not noticed, we set $N = 2^{20}$, $t_{\text{max}} = 32$, $n = 3$, $b = 3$ and $p = 4$.

In the rule 40, the system quickly enters the absorbing state where $x_t = 0$. Figure 2 shows the decay of the residual error as a function of the training span. The average of $\log_{10} E$ is taken over the threads of parallel computation as well as the standard deviations that are indicated by error bars (the values of the error larger than the median are not used in this statistical evaluation). In the panel (a) for $b = 2$, we observe the tendency that both of the error and the standard deviation of decreases with the training span. The former approach a certain destination value, which decreases with $n$ and almost saturates for $n \geq 2$. This is because the number of the independent variable equals $K - 2 = 2$. In the panel (b) for $b = 3$, the destination value keeps decreasing up to $n = 4$. The residual error for $b = 4$ is slightly larger than that for $b = 3$ (not shown). It is reasonable that $b = 3$ yields a good result because the ECA has three-body interaction.

Figure 3 shows the $N$-dependence of the residual error. We observe the tendency that the error decrease with $N$ and converge to a certain value for $n \leq 2$. On the other hand, the error keeps decreasing as $N^{-0.87}$ up to $N = 2^{20}$ for $n \geq 3$, which does not conflict with the existence of an exactly closed DS in the thermodynamic limit.

Figure 4 shows the time evolution of the one component of $X$ for several initial states, which is calculated by using the best result of $(w, W)$. The symbols denotes the solutions of the microscopic DS, where $x$ is transformed to $X$ at each time. The solid lines are the
FIG. 3. The residual error as a function of the system size of CA.

FIG. 4. Time evolution of the one component of X for $2 \leq t \leq t_{\text{max}}$ for the Rule 40. Time is shifted arbitrarily in order to avoid the overlaps among the results of the individual samples.

solutions of the macroscopic DS starting with the common value of X at $t = 1$. These agree well and the relaxation seems to be described by an exponential function.

In the rule 108, the system quickly enters a periodic orbit where $x_{t+2} = x_t$. Figure 5 shows the relation between $X_{1t}$ and $X_{1,t+1}$ for $n = 1$ that are calculated by using the best w. The panel (b) with $b = 3$ exhibits $X_{1,t+1} = X_{1t}$ with high precision, which means that $X_{1t}$ is a conserved quantity. Note that this is conserved even before the system enter the periodic orbit. In the panel (a) with $b = 2$, $X_{1,t+1}$ does not look like a single-valued function of $X_{1t}$; the data for small $t$ are deviated on the upper side. Again, $b = 2$ is insufficient to close the DS.

The Rule 110 shows the behavior at the edge of chaos. The system relax to the steady state where periodic and disordered behaviors are mixed both in time and space. As shown in Fig. 6 the time evolution of X is nearly monotonic relaxation to a unique steady state.
FIG. 5. The relation between $X_{1t}$ and $X_{1,t+1}$ with $n = 1$ in the rule 108. The block size is $b = 1$ in (a) and $b = 3$ in (b). The symbols and the solid lines indicates $X_{1,t+1}$ of 4096 samples and the regression curves $F_1$, respectively.

FIG. 6. Time evolution of the one component of $\mathbf{X}$ for $2 \leq t \leq t_{\text{max}}$ for the Rule 110.

as well as in the rule 40. The residual error is larger than that for the Rule 40 due to fluctuation.

The Rule 90 shows fully chaotic behavior. In Fig. 7 the solutions of microscopic and macroscopic DSs seriously disagree. Remarkably, the best residual error is $10^{-3.28}$, which is less than $10^{-3.12}$ for the rule 110. Actually, the solutions of the two DSs agree well in short time but the difference increases with time. This reminds us the orbital instability of chaotic dynamics. It is observed that $X_{1t}$ jumps up and down irregularly and the value is nearly discrete.

In this paper, we propose a general framework of the dimensional reduction of the DSs by machine learning and show the application to the ECA. The obtained macroscopic DSs actually exhibit the plausible features. This method is expected to work also for stochastic
FIG. 7. Time evolution of the one component of $X$ for $2 \leq t \leq t_{\text{max}}$ for the Rule 90.

processes, such as the contact process and the Glauber dynamics, as far as we treat self-
averaging quantities. It is challenging to develop feasible implementations for systems on off-
lattice and systems with continuous degrees of freedom. There are some extensional usages of
the present method: searching a conserved quantity by fixing $n = 1$ and $F(X) = X$, and
making prediction formula for a quantity of interest, e.g. internal energy, by fixing one
component of $X$.

This work was supported by JSPS KAKENHI Grant Number JP18K03469.

**The implementation of training:** Before the training, we make $2^{16}$ samples of $(\bar{X}(x_t), \bar{X}(x_{t+1}))$
by the independent numerical simulations of the microscopic DS. We subtract a constant
vector from both $\bar{X}(x_t)$ and $\bar{X}(x_{t+1})$ so that $\bar{X}(x_t) = 0$, which also yields $\bar{X}(x_{t+1}) = 0$.
By using these data, we perform the following computation almost independently in 32
threads. First, we initialize the all components of $(w, W)$ to be uniform random numbers
in $[-0.5, 0.5])$. Next, we randomly choose the size of minibatch $S_{\text{mb}}$ from $\{2^9, 2^{10}, \ldots, 2^{16}\}$
and divide the whole data into $2^{16}/S_{\text{mb}}$ minibatches. Then, we continue updating $w$ by
the Mini-batch gradient descent with ADAM [8]. At each step, we randomly choose the
minibatch to use and normalize $X$ by scaling $w$. Every 16 updates of $w$, we renew $W$ by
solving $\bar{F}(X(x_t))^t \bar{F}(X(x_t)) W = \bar{F}(X(x_t))^t X(x_{t+1})$, where $t$ on the left shoulder denotes
the transposition. Every $2^{17}/S_{\text{mb}}$ updates of $w$, we evaluate the error by using the test
data consists of $2^{16}$ samples. We store the record-low error and $(w, W)$ in each thread, and
replace them if a lower error is obtained. Since the test data is different from the training
data, we don’t underestimate the residual error due to overfitting. We set $2^{20}/S_{\text{mb}}$ updates
of $w$ a unit of training, whose time cost rarely depends on $S_{\text{mb}}$. At the end of each unit,
we rank the the momentary errors of the threads and initialize w and W is in the threads whose momentary error is larger than the median. It is likely that these threads are trapped in metastable states. This operation is the only exception of the independence among the threads. After that, we make $S_{mb}$ half or twice randomly as far as $2^9 \leq S_{mb} \leq 2^{16}$. Before entering the next unit, we renew both of the training data and the test data. Strictly speaking, these data are made by randomly choosing $2^{16}$ samples from the total stock of $2^{18}$ samples and shuffling the order.

* nogawa@med.toho-u.ac.jp

[1] E. N. Lorentz, J. Atmos. Sci. 20, 130 (1963).

[2] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, Rev. Mod. Phys. 91, 045002 (2019).

[3] M. Schmidt and H. Lipson, Science 324, 81 (2009).

[4] S. L. Brunton, J. L. Proctor, and J. N. Kutz, 113, 3932 (2016).

[5] K. Harada, Phys. Rev. E 84, 056704 (2011).

[6] K. Harada, Phys. Rev. E 92, 012106 (2015).

[7] S. Wolfram, Rev. Mod. Phys. 55, 601 (1983).

[8] D. P. Kingma and J. Ba, in 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, edited by Y. Bengio and Y. LeCun (2015).