Interpretable Conservation Law Estimation by Deriving the Symmetries of Dynamics from Trained Deep Neural Networks

本武 陽一　　統計的機械学習研究センター　特任助教

●Abstraction

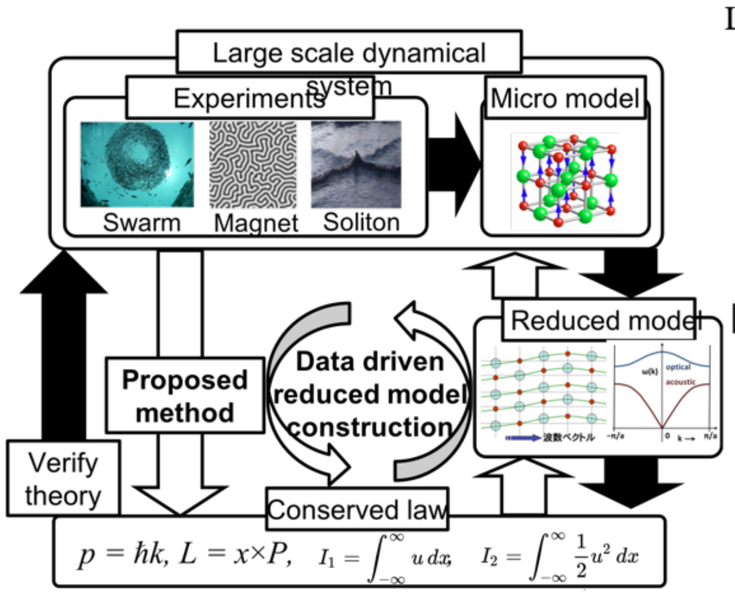
We propose a framework to connect the deep neural networks (DNNs) to a study on conservation laws, the fundamental topic in physics. Our framework does not aspire to conduct physical data analysis using DNNs, but rather to find interpretable physical information from trained DNNs. Using Noether's theorem and an efficient sampling method, the inference of conservation laws was achieved through extracting symmetries from trained DNNs. It applies to a wide range of phenomena, including symmetries, because it only uses the general properties of DNNs and Hamiltonian mechanics. It will support physicists in constructing a minimal model of complex systems.

Paper: https://arxiv.org/abs/2001.00111

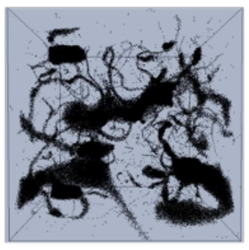
Video: https://www.youtube.com/user/MrKeaton2012

●Introduction

Reduced model and conserved law

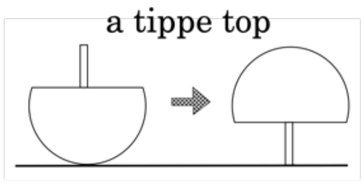


Large scale swarming behavior



[Mototake and Ikegami 15, 16]

Conservation law of a tippe top



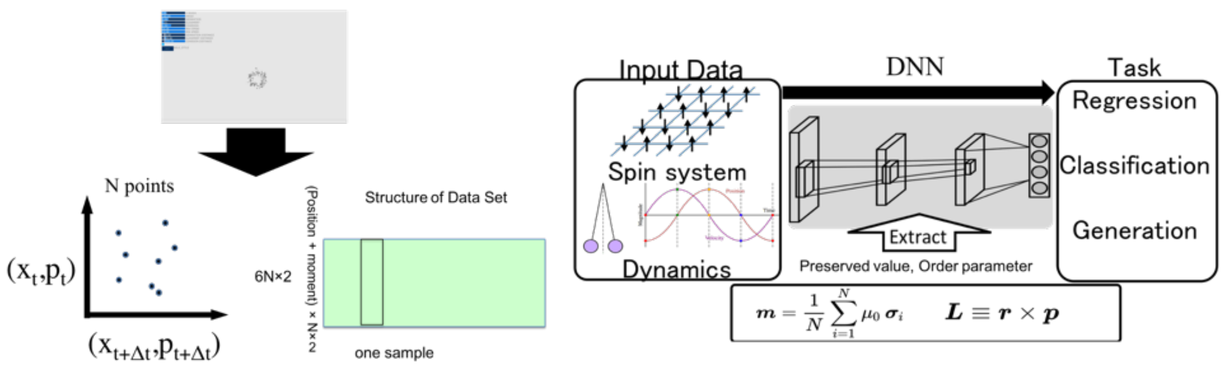
[Moffatt and Shimomura 02]

Noether's theorem

Noether's theorem connects the continuous symmetry of the Hamiltonian system and its conservation law. Considering the Hamiltonian systems in 2-d dimensional phase space $(\boldsymbol{q}, \boldsymbol{p})$, let the system's Hamiltonian be $H(\boldsymbol{q}, \boldsymbol{p})$. Assuming that the Hamiltonian $H(\boldsymbol{q}, \boldsymbol{p})$ and the canonical equations (equations of motion), $\frac{\partial H(\boldsymbol{q},\boldsymbol{p})}{\partial \boldsymbol{q}} = -\dot{\boldsymbol{p}}$ and $\frac{\partial H(\boldsymbol{q},\boldsymbol{p})}{\partial \boldsymbol{p}} = \dot{\boldsymbol{q}}$, are invariant for infinitesimal transformation, $(t', q_i', p_i') = (t + \delta t, q_i + \delta q_i, p_i + \delta p_i)$, where $i = 1 \sim d$. Then, based on Noether's theorem, the conserved value G satisfies the following equation:

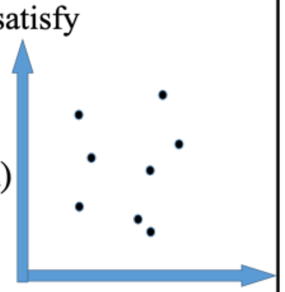
$$(\delta q_j, \delta p_j) = \left(\frac{\partial G_\delta}{\partial p_j}, -\frac{\partial G_\delta}{\partial q_j}\right). \tag{1}$$

Data manifold of time series data



The candidate of invariant transformation of Hamiltonian system must be in transformations $(q,p) \longmapsto (\mathbf{Q}, \mathbf{P}) := (\mathbf{Q}(q,p), \mathbf{P}(q,p))$ which satisfy below condition:

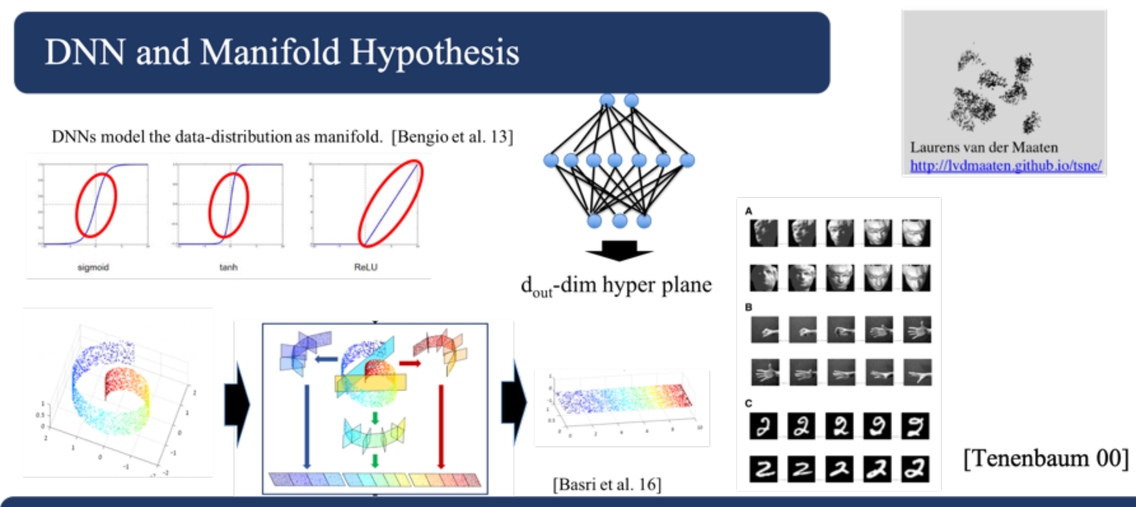
$$\left\{ q_{t+\Delta t}, p_{t+\Delta t}, q_t, p_t \;\middle|\; H(q_t,p_t) = E_i, p_{t+\Delta t} = p_t - \frac{\partial H(q_t,p_t)}{\partial q_t}, q_{t+\Delta t} = q_t + \frac{\partial H(q_t,p_t)}{\partial p_t} \right\}$$
$$= \left\{ Q_{T+\Delta T}, P_{T+\Delta T}, Q_T, P_T \;\middle|\; H(q_t,p_t) = E_i, p_{t+\Delta t} = p_t - \frac{\partial H(q_t,p_t)}{\partial q_t}, q_{t+\Delta t} = q_t + \frac{\partial H(q_t,p_t)}{\partial p_t} \right\}.$$



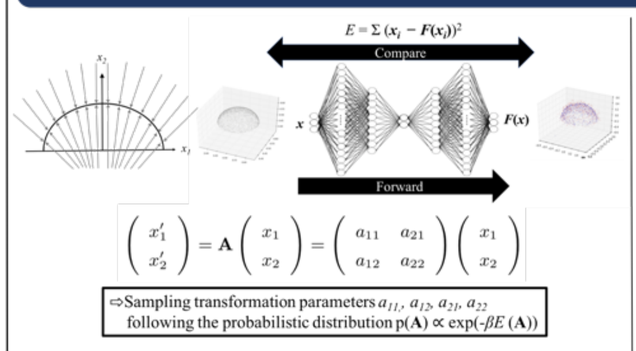
➡It's possible to estimate the conservation law directly from the time-series dataset!!

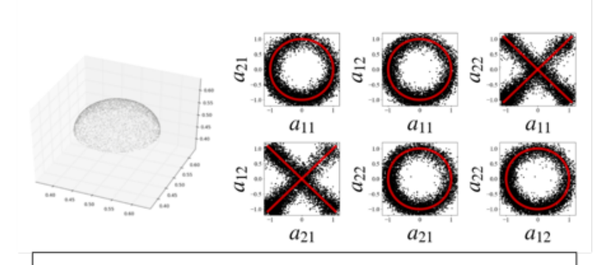
●Method

DNN and Manifold Hypothesis

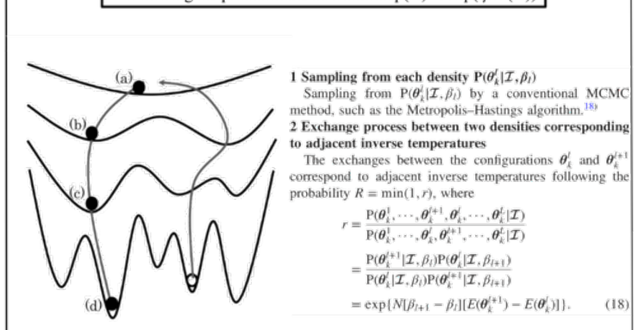


Method to estimate symmetry of manifold structure of time series data





$$\hat{h}_k(c_k, b_1, b_2, \cdots, b_{d_\theta'}; \beta, \gamma, d_\theta') := \sum_{s_0=0}^{d_b}\sum_{s_1=0}^{d_b}\cdots\sum_{s_{d_\theta'}=0}^{d_b} \gamma_{s_0 s_1 s_2 \cdots s_{d_\theta'}} \beta_{s_0 s_1 s_2 \cdots s_{d_\theta'}} c_k^{s_0} b_1^{s_1} b_2^{s_2} \cdots b_{d_\theta'}^{s_{d_\theta'}} = 0,$$



1 Sampling from each density $\mathrm{P}(\theta_k^i|\mathcal{I}, \beta_i)$
Sampling from $\mathrm{P}(\theta_k^i|\mathcal{I}, \beta_i)$ by a conventional MCMC method, such as the Metropolis–Hastings algorithm.[18]

2 Exchange process between two densities corresponding to adjacent inverse temperatures
The exchanges between the configurations θ_k^i and θ_k^{i+1} correspond to adjacent inverse temperatures following the probability $R = \min(1, r)$, where

$$r = \frac{\mathrm{P}(\theta_k^1, \cdots, \theta_k^{i+1}, \theta_k^i, \cdots, \theta_k^I|\mathcal{I})}{\mathrm{P}(\theta_k^1, \cdots, \theta_k^i, \theta_k^{i+1}, \cdots, \theta_k^I|\mathcal{I})} = \frac{\mathrm{P}(\theta_k^{i+1}|\mathcal{I}, \beta_i)\mathrm{P}(\theta_k^i|\mathcal{I}, \beta_{i+1})}{\mathrm{P}(\theta_k^i|\mathcal{I}, \beta_i)\mathrm{P}(\theta_k^{i+1}|\mathcal{I}, \beta_{i+1})} = \exp\{N[\beta_{i+1} - \beta_i][E(\theta_k^{i+1}) - E(\theta_k^i)]\}. \tag{18}$$

Algorithm 1 Estimation of the invariant transformation set

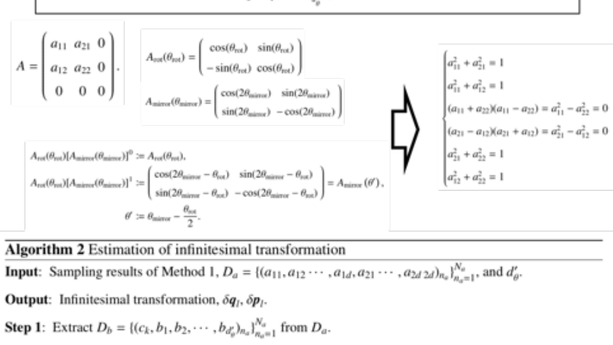
Input: dataset $D = \left\{\boldsymbol{q}_{t_i}^i, \boldsymbol{p}_{t_i}^i, \boldsymbol{q}_{t_i+\Delta t}^i, \boldsymbol{p}_{t_i+\Delta t}^i\right\}_{i=1}^N$ in a given coordinate system.

Output: Invariant transformation set $D_a = \{(a_{11}, a_{12} \cdots, a_{1d}, a_{21} \cdots, a_{2d\,2d})_{n_a}\}_{n_a=1}^{N_a}$.

Step 1: Train the deep autoencoder with dataset D.

Step 2: Using the trained deep autoencoder and REMC method, sampling transformation parameters $a_{11}, a_{12}, a_{21}, \cdots, a_{2d\,2d}$ from multiple probability distributions $P'(a_{11}, a_{12}, a_{21}, \cdots, a_{2d\,2d})$ corresponding to different noise intensities σ'.

Step 3: Select σ' from the distribution structure of the sampling results and output the sampling result of the selected σ' state as D_a.



Algorithm 2 Estimation of infinitesimal transformation

Input: Sampling results of Method 1, $D_a = \{(a_{11}, a_{12} \cdots, a_{1d}, a_{21} \cdots, a_{2d\,2d})_{n_a}\}_{n_a=1}^{N_a}$, and d_θ'.

Output: Infinitesimal transformation, $\delta \boldsymbol{q}_l$, $\delta \boldsymbol{p}_l$.

Step 1: Extract $D_b = \{(c_k, b_1, b_2, \cdots, b_{d_\theta'})_{n_a}\}_{n_a=1}^{N_a}$ from D_a.

Step 2: Fit D_b with the implicit polynomial function $\hat{h}_k(c_k, b_1', b_2', \cdots, b_{d_\theta'}'; \beta, \gamma, d_\theta')$ [Eq. (54)] for each c_k.

Step 3: Estimate the likelihood [Eq. (G1)] by numerical integration of Z [Eq. (G2)].

Step 4: Select the indicator vector γ and the dimension d_θ' of $M_{\text{invariant}}$ in Eq. (54) for each c_k using the BIC.

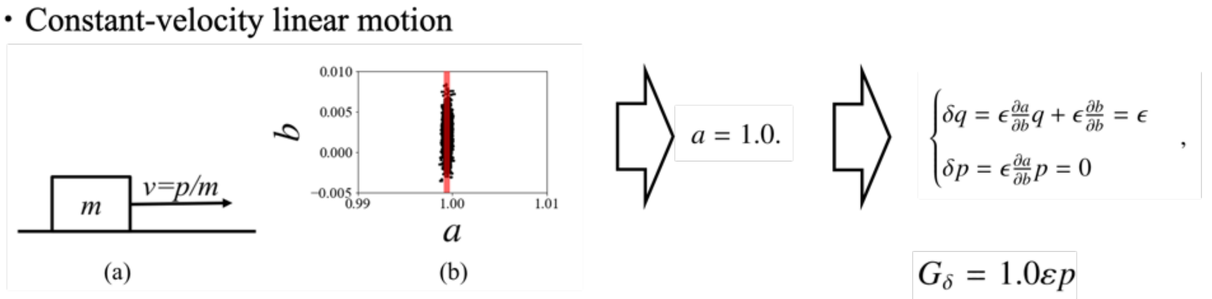
Step 5: Determine whether the Jacobi matrix $J_{kl} = \frac{\partial \hat{h}_k(c_k, b_1, \cdots, b_{d_\theta'})}{\partial b_l}$ is nonsingular. If J_{kl} is singular, return to Step 1 and re-extract D_b'.

Step 6: Differentiate the obtained simultaneous equations with respect to b_l around a point e_I to obtain Eq. (52).

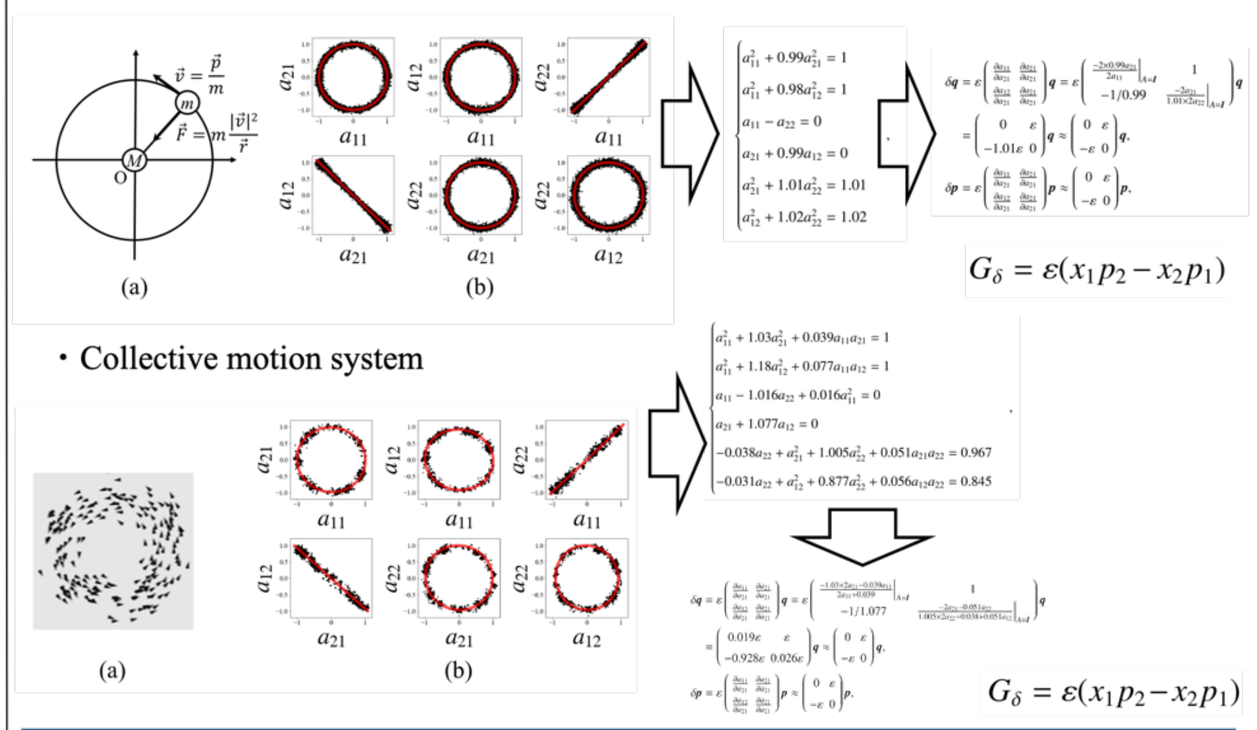
Step 7: Solve the simultaneous equations in Eq. (52) and obtain the infinitesimal transformation, $\delta \boldsymbol{q}_l$, $\delta \boldsymbol{p}_l$.

●Results

- Constant-velocity linear motion



- Two-dimensional central force system



- Collective motion system

●Summary and Discussion

Targeting the collective motion of living things, we show that our framework will be a powerful tool for physicists constructing a reduced model of complex systems. We believe that our work greatly contributes to the construction of a reduced model of complex systems such as a collective motion system