1I4-J-2-01

DNNを用いて抽出された力学系の対称性からの保存量推定

Conservation-Law Estimation from Symmetric Property of Dynamical System Extracted using DNN

本武　陽一
Mototake Yoh-ichi

東京大学大学院新領域創成科学研究科
Graduate School of Arts and Science, The University of Tokyo

It is suggested that Deep Neural Networks (DNN), which continues to develop in recent years, has a function to extract information of data sets necessary to achieve a given task by modeling the distribution as a manifold. In addition to confirming the usefulness of DNN technology, numerous researchers and engineers are developing various DNN algorithms and tuning parameters. This situation means that enormous knowledge on the manifold structure for various data sets is being accumulated. The purpose of this research is to propose a method to extract manifold structure with complex shape extracted in an interpretable form. Specifically, we propose a method to extract the symmetry of manifold for coordinate transformation. Based on the Noether's theorem in physics, we develop the method to estimate the conservation law of the system. Applying the proposed method to the time series data of the moving object according to the central force potential, it was confirmed that symmetry according to the conservation law of angular momentum could be extracted.

1. 導入

近年発達を続ける Deep Neural Networks（以下，DNN）が，与えられたタスクを達成するために必要なデータセットの情報を，その分布を多様体としてモデル化することで抽出する機能を持つことが示唆されている [1][2][3][4][5][6][7][8]．このような視点に立つと，DNN の写像関数は多様体を定義するものと理解される．しかしながら，この写像関数は一般に，非常に多数のパラメータによって構成された非線形関数となるため，そこからデータセット分布の持つ情報を解釈可能な形で抽出することは困難である．

これに対して我々は，この DNN がモデル化した多様体構造についての情報を，その入力空間の連続的な変換に対する対称性として抽出する手法を提案する．これによって，複雑な写像関数を対称性という指標でより単純に評価することが可能となる．

対称性から対象のデータセットの理解を行うためには，対称性とデータセットの性質との間を結ぶ理論が必要となる．物理学の分野では，この対称性と対象の性質に関する理論の構築が長年行われて来た．ネーターの定理はその代表例で，ハミルトニアンの座標変換に対する普遍性と，系の保存量とを結びつけることが可能となる．本研究では，提案手法を中心力ポテンシャル系の時系列データに適用し，そこからハミルトニアンの対称性を抽出することを試みた．この結果，角運動量に対応する対称性が抽出されることを示唆する結果が得られた．

2. 手法

2.1 DNN と多様体

多様体は，ユークリッド空間を貼り合わせることによって構成される空間である。大まかな例としては，地図と二次元の多様体である地球表面がある．我々は，地球表面を 2 次元のユークリッド空間である地図のラミネーションとして理解する．より正確には，d_m 次元微分可能多様体 M とは，多様体上の任意の点 $\boldsymbol{x}^0$ で接空間 $T_{\boldsymbol{x}^0}M$ と呼ばれる d 次元ユークリッド空間と同相な近傍を持つ位相空間である．本研究では，DNN はデータセットの分布を，多様体として近似的にモデル化すると考える．その上で，DNN はデータ多様体をそれと同じ次元のユークリッド空間に近しい大域的座標系へ写像する機能を持つと考える．本研究では，データ分布を多様体として近似した際に現れる構造をデータ多様体と呼ぶ．

微分同相写像は，d_m 次元多様体 M を d_m 次元多様体 N へ写像する微分写像関数として定義される．逆関数定理より，微分写像関数のうち，そのヤコビアン行列が多様体 M 上の全ての点で正則行列となるものは微分同相写像となる．つまり，ある微分写像関数のヤコビアン行列が入力空間と出力空間上の全ての点で正則行列となる場合，入力空間と出力空間は多様体としてモデル化されているということができる．多様体仮説で取り扱うのは，ユークリッド空間中の多様体となる．この場合，多様体を定義する微分写像に求められる条件は，多様体に対応する部分空間から多様体に対応する部分空間への写像に関して同相写像となることである．これは前の議論と同様に，ある微分写像が与えられた場合に，その部分空間上の全ての点に関する写像関数のヤコビアンが正則行列となることと同値となる。つまり、ある部分空間上の全ての点での写像関数のヤコビアン行列の rank が d_m となる場合，その部分空間が多様体としてモデル化されているといえる．これに基づき，我々は DNN 写像関数がモデル化した入出力空間での多様体の接空間を以下の手続きで算出した．今、入力が d_{in} 次元，出力が d_{out} 次元である，L 層 DNN を考える．この DNN の写像関数 $\boldsymbol{F}(\boldsymbol{x}) = (F_1(\boldsymbol{x}), F_2(\boldsymbol{x}), \cdots, F_{d_{out}}(\boldsymbol{x}))$ は、

$$\begin{aligned}\boldsymbol{F}(\mathbf{x}) = \boldsymbol{h}^L &= \boldsymbol{f}(\boldsymbol{w}^{L-1}\boldsymbol{h}^{L-1} + \boldsymbol{b}^{L-1}) \\ &= \boldsymbol{f}(\boldsymbol{w}^{L-1}\boldsymbol{f}(\cdots \boldsymbol{f}(\boldsymbol{w}^1\boldsymbol{x} + \boldsymbol{b}^1)\cdots) + \boldsymbol{b}^{l-1})\end{aligned} \tag{1}$$

によって与えられる．ここで，$\boldsymbol{h}^l = (h^l_0, h^l_1, \cdots, h^l_{d_l})$ は，d_l

連絡先: 本武陽一，東京大学大学院総合文化研究科，mototake@sacral.c.u-tokyo.ac.jp

次元の第 l 層の出力とし，$\boldsymbol{f}(\cdot)$ を，

$$\begin{aligned}\boldsymbol{f}(\boldsymbol{w}^{l-1}\boldsymbol{h}^{l-1}+\boldsymbol{b}^{l-1}) &= (f_1, f_2, \cdots, f_{d_l}) \\ f_j &= f\left(\sum_i^{d_{l-1}}\left(w_{ij}^{l-1}h_i^{l-1}+b_j^{l-1}\right)\right)\end{aligned} \tag{2}$$

とした．f は活性化関数と呼ばれ，良くシグモイド関数や LeRU 関数等が用いられる．従って，ある入力 $\boldsymbol{p}$ が与えられた場合の DNN 写像関数のヤコビアン行列 $\boldsymbol{J}$ は，

$$\begin{aligned}J_{ij}^{\mathrm{DNN}} &= \frac{\partial F_j(\boldsymbol{x})}{\partial x_i} \\ &= \sum_{<k_{l-1},k_{l-2},\cdots,k_1>} \frac{\partial h_j^L}{\partial h_{k_{l-1}}^{l-1}}\frac{\partial h_{k_{l-1}}^{l-1}}{\partial h_{k_{l-2}}^{l-2}}\cdots\frac{\partial h_{k_1}^1}{\partial x_i}\end{aligned} \tag{3}$$

で与えられる．また，各層間の写像関数のヤコビアンを

$$\boldsymbol{J}^l = \begin{pmatrix} \frac{\partial h_1^l}{\partial h_1^{l-1}} & \cdots & \frac{\partial h_1^l}{\partial h_{d_{l-1}}^{l-1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial h_{d_l}^l}{\partial h_1^{l-1}} & \cdots & \frac{\partial h_{d_l}^l}{\partial h_{d_{l-1}}^{l-1}} \end{pmatrix} \tag{4}$$

とした場合に，全体のヤコビアンは

$$\boldsymbol{J}^{\mathrm{DNN}} = \boldsymbol{J}^L\boldsymbol{J}^{L-1}\cdots\boldsymbol{J}^1 \tag{5}$$

で与えられる．例えば活性化関数 f としてシグモイド関数を用いた場合，

$$\boldsymbol{J}^l = \begin{pmatrix} h_1^{l-1}(1-h_1^l) & \cdots & h_{d_{l-1}}^{l-1}(1-h_1^l) \\ \vdots & \ddots & \vdots \\ h_1^{l-1}(1-h_{d_l}^l) & \cdots & h_{d_{l-1}}^{l-1}(1-h_{d_l}^l) \end{pmatrix} \tag{6}$$

となる．このように写像関数から定義された多様体上のある点 $\boldsymbol{p}$ での局所近傍座標系である接空間 $T_{\boldsymbol{p}}M$ は，点 $\boldsymbol{p}$ 周りのでヤコビアン行列 $\boldsymbol{J}$ の縮退していない基底空間であるユークリッド空間として与えられる．

この描像は，DNN がシグモイド関数のほぼ線形な部分をつなぎ合わせた曲がった空間上で主成分分析のようなことを行うものとして理解される．その際圧縮される方向は，多様体から遠く離れた点まで及ぶことに注意されたい（図 2）．また，シグモイド関数の線形部分が有限であることから，多様体から離れるほど，圧縮領域の隙間となる空間が拡大する．このような空間は，シグモイド関数の性質から，0 あるいは 1 に吸引されていくと考えられる．

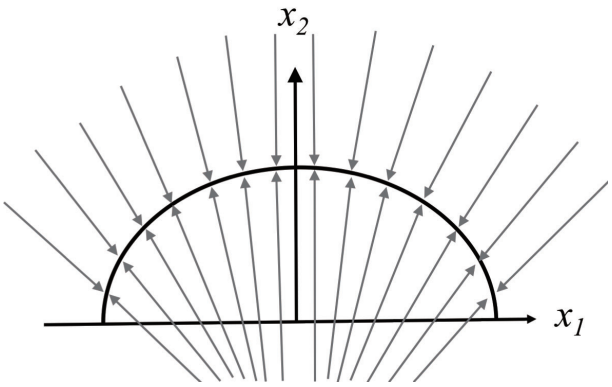

図 1: DNN におけるデータ分布の多様体への引き込みメカニズムの模式図

2.2 DNN のモデル化した多様体からの対称性抽出手法

前節の議論より，入力空間で多様体上にないデータ点は，多様体上に吸引されるとわかる．入力空間から見た中間層でのデータ点の移動は図 2 のような模式図に従うと考えられる．中間層において一度多様体上に吸引されたデータ点は，それよりも後の層において多様体上のデータと分離されることはない．したがって，出力層 $\boldsymbol{F}(\mathbf{x})$ において，全てのデータは多様体上，あるいは空間の角となる 0 と 1 の組み合わせで与えられる座標点に縮退する．

これをふまえると，あるデータ多様体を学習した DNN に対して，新たに与えられたデータセット $\{\boldsymbol{x}_i\}_{i=1}^N$ がそのデータ多様体上にあるかは，それが出力空間に写像された際のデータ集合 $\{\boldsymbol{F}(\boldsymbol{x}_i)\}_{i=1}^N$ との間の二乗誤差

$$E(\{\boldsymbol{x}_i\}_{i=1}^N; \boldsymbol{F}(\boldsymbol{x})) = \sum_{i=1}^N [\boldsymbol{x}_i - \boldsymbol{F}(\boldsymbol{x}_i)]^2 \tag{7}$$

によって評価されるとわかる（図 2）．

なぜならば，入力空間で多様体から離れたデータほど，出力空間で多様体上に落ち込むまでに移動した距離が遠く，結果として二乗誤差が大きくなるためである．

今知りたいのは，DNN の訓練時に用いたデータ多様体分布がどのような座標変換 $\mathbf{A}$ に対して普遍となるかということである．先ほど議論した多様体からの距離を測る指標 $E(\{\boldsymbol{x}_i\}_{i=1}^N; \boldsymbol{F}(\boldsymbol{x}))$ を用いると，データ多様体の分布構造を変えない座標変換は，

$$E(\{\mathbf{A}\boldsymbol{x}_i\}_{i=1}^N; \boldsymbol{F}(\mathbf{A}\boldsymbol{x})) = \sum_{i=1}^N [\mathbf{A}\boldsymbol{x}_i - \boldsymbol{F}(\mathbf{A}\boldsymbol{x}_i)]^2 \tag{8}$$

が小さな値となる $\mathbf{A}$ を集めてくれば良いとわかる．

具体的には，行列 $\mathbf{A}$ の $d_{in}\times d_{in}$ 個ある要素 a_{jk} についてサンプリングを行うことで，$E(\{\mathbf{A}\boldsymbol{x}_i\}_{i=1}^N; \boldsymbol{F}(\mathbf{A}\boldsymbol{x}))$ がある閾値以下となる $\mathbf{A}$ の集合が得られる．

2.3 ネーターの定理

系の対称性と保存量との関係について，ネーターによって証明された定理である [9]．ここで，d 次元一般化座標を，$\mathbf{q}(t)$ と $\mathbf{p}(t)$ とし，系のハミルトニアンを $H(q,p;t)$ とする．今，ハミルトニアン $H(q,p;t)$ が，ある微小な座標変換

$$\begin{aligned}t' &= t+\delta t \\ q_j' &= q_j+\delta q_j\end{aligned} \tag{9}$$

（ただし，j=1 ∼ d）に対して普遍であるとする．すると，ネーターの定理より微小変換の生成子 G_δ が時間によらず普遍な量となる．つまり，保存量となる．ここで微小変換の生成子 G_δ は，

$$(\delta q_j, \delta p_j) = \left(\frac{\partial G_\delta}{\partial q_j}, \frac{\partial G_\delta}{\partial p_j}\right) \tag{10}$$

を満たすものとして定義される．

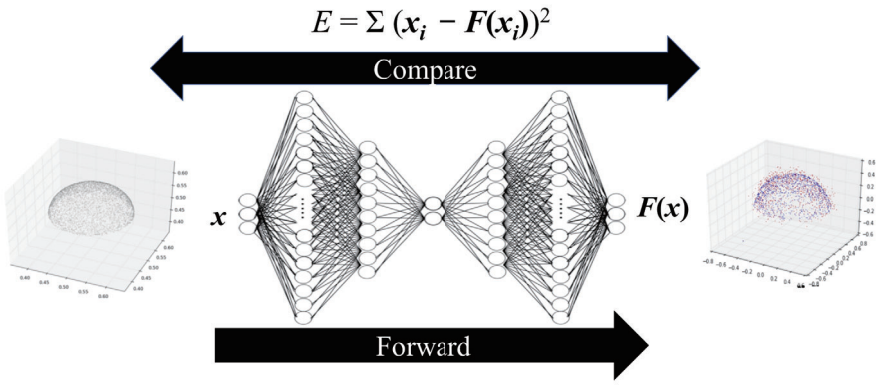

図 2: 対称性の抽出．多様体上からの距離を，入力データと出力データの 2 乗誤差から定量化．

2.4 時系列データの対称性とネーターの定理

次に，ネーターの定理を適用するために必要なハミルトニアン $H(q,p;t)$ の微小変換に対する普遍性を，位相空間での時系列データ $\{\mathbf{q}^i(t),\mathbf{p}^i\}_{i=1}^N$ から得る方法について述べる．ハミルトニアン力学系において，力学系の時間発展方程式は，$\mathbf{q},\mathbf{p}$ についての一階微分方程式となる．このことから，一つの仮定を設定した．それは，「与えられたハミルトニアン $H(q,p;t)$ の元である一定エネルギー E で運動を行う系が，$4d$ 次元空間 $(\mathbf{q}(t),\mathbf{p}(t),\mathbf{q}(t+1),\mathbf{p}(t+1))$ に取り得る全ての状態が持つ等エネルギー面（多様体 M_E）の持つ対称性が，ハミルトニアンの対称性と関係する.」というものである．これはすなわち，あらゆる初期値から生成した同じエネルギーを持つ運動の時系列データの持つ多様体の対称性を抽出すれば，保存量の候補が得られるという仮定となっている．この仮定によって得られた結果の妥当性は，最終的に得られた保存量が時間普遍であることを確認することで容易に確認される．この M_E が持つ対称性を 2.2 節の方法で推定することで，保存量推定を行うのが本提案手法の枠組みである．

2.5 変換の生成子

最後に，2.2 節の方法で得られるハミルトニアン普遍な大域的変換 $\mathbf{A}$ から，微小変換の生成子 G_δ を取り出す方法を，3 次元回転の行列と SO(3) の生成子を例にとって説明する．

Z 軸まわりの回転行列を書き下すと，

$$A_z = \begin{pmatrix} cos(\theta_z) & sin(\theta_z) & 0 \\ -sin(\theta_z) & cos(\theta_z) & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{11}$$

となる．今，この回転角 θ_z が，非常に小さい（$=\delta\theta_z$）として，この行列を θ_z についてテーラー展開し，一次の項までをとると，

$$A_z \sim \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} + \begin{pmatrix} 0 & -\delta\theta_z & 0 \\ \delta\theta_z & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{12}$$

となる．この右辺第二項が Z 軸周りでの回転の生成子となる．他の軸についても同様に生成子が得られる．

実際にサンプリングして得られる $\mathbf{A_z}$ においては，隠れ変数である θ_z による変換行列のパラメトライズはできないため．要素 a_{ij} の間の関係の回帰や，それらの分布の分析などから，

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{13}$$

周りでの微小変動の方向を見積もることで，回転の生成子を得た．

3. 結果と考察

物理学データへの手法の適用を行う．具体的に適用するのは，以下のハミルトニアン

$$H(\mathbf{x},\mathbf{p}) = \frac{\mathbf{p}^2}{2m} - \frac{k}{r} \tag{14}$$

に従う 2 次元空間中の中心力運動である．実験では，$k = 10$,$m = 1$ とし，さらに，運動を円運動に限定した上で，系の取りうる全ての状態によってトレーニングデータを生成した．

サンプリング対象となる変換行列は，配位空間 $\mathbf{x} = (x_1, x_2)$ を，

$$\begin{pmatrix} x_1' \\ x_2' \end{pmatrix} = \mathbf{A}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \tag{15}$$

と変換する．この時，運動量空間 $\mathbf{p} = (p_1, p_2)$ は，

$$\begin{aligned} \begin{pmatrix} p_1' \\ p_2' \end{pmatrix} &= \begin{pmatrix} m\dot{x'}_1 \\ m\dot{x'}_2 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{pmatrix}\begin{pmatrix} m\dot{x}_1 \\ m\dot{x}_2 \end{pmatrix} \\ &= \begin{pmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{pmatrix}\begin{pmatrix} p_1 \\ p_2 \end{pmatrix} \end{aligned} \tag{16}$$

と変換される．したがって，この 4 つの変換行列の要素についてサンプリングを行う．ちなみに今回は，時間変換に対する対称性は，明らかであるので検討対象としなかった．ちなみに，時間方向の対称性は，エネルギー保存則を導く．

ここではメトロポリスサンプリングを用いた．メトロポリスサンプリングでは，全てのパラメータについて step 数は 0.2 と設定され，分散 0.2 のガウスノイズを仮定した．その上で，1,000step の burn in の後に 9,000step のサンプリングを実行した．この結果，図 3 のような結果が得られた．

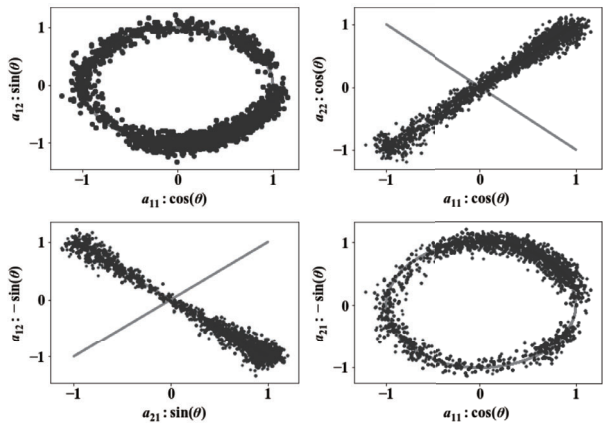

図 3: 中心力ポテンシャル中を運動する物体の時系列データセットに対して，提案手法によって得られたデータ多様体の分布を変えない対称性変換 $\mathbf{A}$ の要素 a_{jk} の比較．$\mathbf{A}$ が回転行列となっている場合に得られるであろう分布構造を黒線で表した．ただし，a_{11}-a_{22} と a_{12}-a_{21} の図における直線の片方は，鏡像変換の従う分布構造についてもプロットした図となっている．

得られた結果より，変換の生成子を推定する．今回得られた結果の図から変換が単位行列となる $a_{11} = a_{22} = 1$ 周りの微小変動を考察すると，a_{11}-a_{12} と a_{11}-a_{21} の関係（図 3）より，生成子の a_{11},a_{22} 成分が 0 となり，a_{12} と a_{21} 成分が $\pm\delta\theta_y$ 方向を持つとわかる．さらに，a_{21}-a_{12} の関係より，生成子の a_{11},a_{22} 成分の符号が逆であることがわかる．以上より，得られた変換の生成子は，

$$\begin{pmatrix} 0 & -\delta\theta_y \\ \delta\theta_y & 0 \end{pmatrix} \tag{17}$$

となる．これはちょうど SO(2) の対称性の生成子となっており，これより，角運動量が保存することがわかる．

以上のように，提案手法によって力学系の時系列データから，そのハミルトニアンの対称性に関する情報を抽出し，保存則が推定されることが確認された．

4. まとめと今後の展望

データ多様体をモデル化した DNN から，その多様体の持つ対称性に関する情報を抽出する手法を構築した．そしてその応用として，中心力ポテンシャル中を運動する力学系の時系列データの作るデータ多様体に提案手法を適用した．その結果，力学系が持つ対称性の抽出と，ネーターの定理を介した保存量の推定を実現した．

本提案手法は，さらに複雑な対称性を持つ系への適用も可能であると考えられる．具体的には，中心力ポテンシャル中の力学系が持つ配位空間とは違う空間における隠れた対称性である，SO(4) の抽出と，対応する保存量であるルンゲ-レンツベクトルの推定等の実現が期待される．

近年の DNN 技術の有用性の確認とともに，多数の研究者・技術者が各種の DNN アルゴリズムの開発やパラメータチューニングを行なっている．この状況は，各種データセットに対する多様体構造についての莫大な知見が蓄積されつつあることを意味する．本提案手法は，このような蓄積された莫大な知見を解析するための手法開発に繋がっていくことが期待される．

参考文献

[1] Irie Bunpei and Kawato Mitsuo. Acquisition of internal representation by multi-layered perceptrons. *The Transactions of the Institute of Electronics, Information and Communication Engineers D*, Vol. 73, No. 8, pp. 1173–1178, 1990.

[2] GE Hinton and RR Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 2006.

[3] Pratik Prabhanjan Brahma, Dapeng Wu, and Yiyuan She. Why deep learning works: A manifold disentanglement perspective. *IEEE transactions on neural networks and learning systems*, Vol. 27, No. 10, pp. 1997–2008, 2016.

[4] Ronen Basri and David Jacobs. Efficient representation of low-dimensional manifolds using deep networks. *arXiv preprint arXiv:1602.04723*, 2016.

[5] Salah Rifai, Yann N Dauphin, Pascal Vincent, Yoshua Bengio, and Xavier Muller. The manifold tangent classifier. pp. 2294–2302, 2011.

[6] 本武陽一, 池上高志. Deep neural networks の力学的・幾何学的解析. 人工知能学会全国大会論文集, Vol. JSAI2016, pp. 1A5OS27c1–1A5OS27c1, 2016.

[7] Yhoichi Mototake and Takashi Ikegami. The dynamics of deep neural networks. *Proceedings of the Twentieth International Symposium on Artificial Life and Robotics*, Vol. 20, , 2015.

[8] Y. Mototake. Geometrical structures embedded in high dimensional data sets and deep learning : Analysis and application to dynamical systems. 2016.

[9] AE Noether. Nachr kgl ges wiss göttingen. *Math. Phys. KI II*, Vol. 235, , 1918.