深層ニューラルネット理論の近況

園田　翔　(理研 AIP 深層学習理論チーム)*

2018年7月31日

1. 導入

深層学習による技術革新は未だとどまるところを知らない。深層学習とは深層ニューラルネット（neural network; NN）を学習させる技術の総称である。深層NNとは，図1に示すように入力層と出力層の間に複数の中間層をもつNNである。2006年に初めて深層学習が登場するまで，深層NNは高々5層程度までしか学習できないと信じられていた。2018年現在，深層NNは10,000層まで学習できるようになった (Xiao et al., 2018)。実は § 4 で詳述する通り，理論的には既に無限層モデルが登場している。

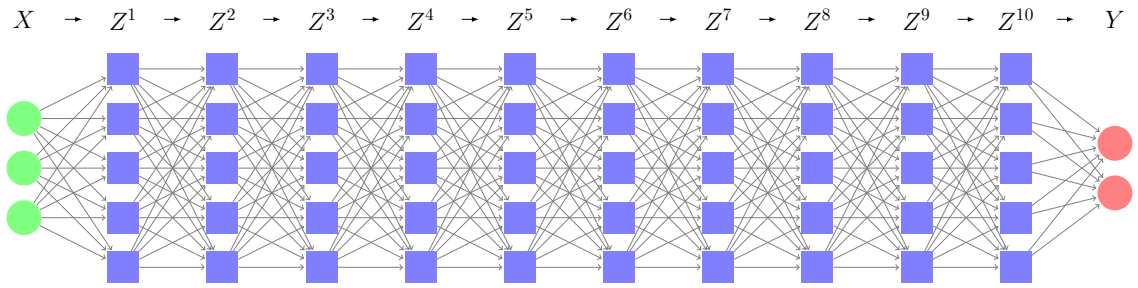

図 1: 深層NNの模式図。各ノードは神経細胞を表し，エッジはシナプス結合（神経発火に伴う信号の入出力関係 $\mathbb{R} \to \mathbb{R}$）を表す。層状に並んだ神経細胞はそれぞれ入力層 X，中間層 Z^l，出力層 Y をなす。信号は入力層から出力層へ一方向に流れるので，深層NNは合成写像 $\mathbb{R}^3 \to \mathbb{R}^5 \to \cdots \to \mathbb{R}^2$ とみなせる。

もちろん，単に深くできることが深層学習の魅力というわけではない。人工知能の要素技術となる様々なタスクにおいて既存手法を圧倒する性能を発揮した実績こそが，真に革新的な技術と評される所以である。深層学習は，2012年に画像認識 (Krizhevsky et al., 2012) や音声認識 (Hinton et al., 2012) といった認識タスクで頭角を表し，今日では自動運転での使用にも耐えうる水準に達しつつある (Redmon et al., 2016)。その後，強化学習では人類最強の棋士を破り (Silver et al., 2016)，画像生成 (Radford et al., 2016) や音声合成 (van den Oord et al., 2016)，説明文生成 (Vinyals et al., 2015) や翻訳 (Wu et al., 2016) といった生成タスクでも存在感を増している。そして，機械学習理論の研究対象としても，深層学習は従来の枠にはまらない型破りの存在であり，新理論の覇権を巡って百家争鳴が続いている。

NNはブラックボックスと呼ばれる。NNのパラメータ空間は，大きいものでは144M次元 (Simonyan and Zisserman, 2015) にも及び，そのうえ学習問題は無数の局所解を

本研究は科研費 (課題番号:18K18113) の助成を受けたものである。
2010 Mathematics Subject Classification: 68T05,35Q62,62G08
キーワード：深層ニューラルネット, 積分表現, リッジレット解析, 輸送理論, Wasserstein 幾何学
* 〒 103–0027 東京都中央区日本橋 1–4–1 理化学研究所　革新知能統合研究センター
e-mail: sho.sonoda@riken.jp
web: https://sites.google.com/view/shosonoda/

もつためである。ブラックボックスのままでは，例えば自動運転のようなシステムに載せる場合に誤動作のリスクを下げきれない。実際，一見うまく学習できているようなNNであっても，入力に摂動を加えると思わぬ誤作動を引き起こせることが知られている。このような入力を敵対的サンプル（adversarial example）という。例えば，バスの画像に雑音を加えてダチョウであると誤認識 (Szegedy et al., 2014) させたり，特殊なメガネをかけた男性を女優と誤認識 (Sharif et al., 2016) させたりできる。NNを含む多くの学習機械を解釈可能（interpretable）ないし説明可能（explanable）なホワイトボックスにする研究は，2016年頃から注目を集めている (Molnar)。

本稿では，1990年代に開発された浅いNNの理論である**積分表現理論**と，近年深層NNの理論として存在感を増しつつある**輸送理論**という二つの観点から，NNのホワイトボックス化に迫る。積分表現理論は，浅いNNを積分表現にして，関数解析ないし調和解析を用いてNNの関数近似能力を調べる理論である。一方，輸送理論は，深層NNを連続力学系として表現して，最適輸送理論やWasserstein幾何学を用いて深層NNの中間写像を調べる理論である。積分表現が表現できるのは中間層が1層の浅いNNに限るが，図1のように中間層どうしを直接合成する形式ではなく，図2のように浅いNNを合成する形式を調べることで，深層NNに対しても積分表現理論が展開できるようになる。

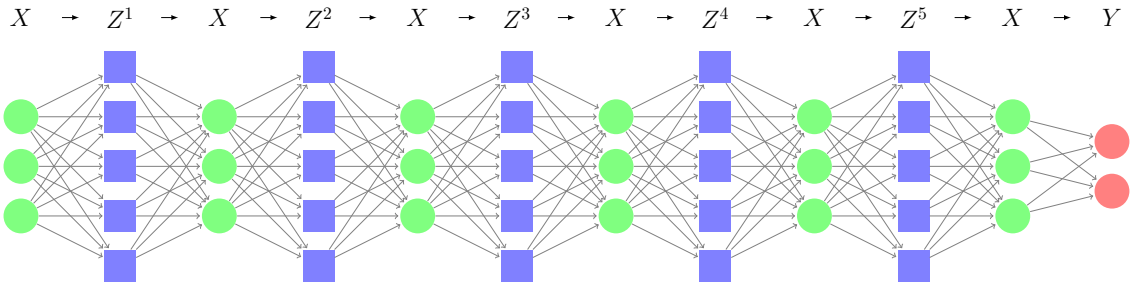

図 2: 浅いNN$X \to Z \to X$を合成する形式の深層NN。浅いNNの理論である積分表現理論と，合成写像の理論である力学系を併用できる。

2. 定式化

2.1. 浅いニューラルネット

次の形式でパラメータ付けられた関数$g : \mathbb{R}^m \to \mathbb{R}$を**浅い** *NN* という

$$g(\boldsymbol{x}; \{(\boldsymbol{a}_j, b_j, c_j)\}_{j=1}^p) := \sum_{j=1}^p c_j \sigma(\boldsymbol{a}_j \cdot \boldsymbol{x} - b_j), \quad \boldsymbol{x} \in \mathbb{R}^m. \tag{1}$$

ただし各jに対し，$(\boldsymbol{a}_j, b_j, c_j) \in \mathbb{R}^m \times \mathbb{R} \times \mathbb{C}$である。$\sigma$は活性化関数と呼ばれる緩増加超関数（$\mathcal{S}'$）である。例えばガウシアン$\exp(-z^2/2)$やシグモイド関数$\tanh z$，rectified linear unit (ReLU) z_+などが典型的である。特に必要のない限り，パラメータは明記せず単に$g(\boldsymbol{x})$と書く。

訓練データを$D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \subset \mathbb{R}^m \times \mathbb{R}$とする。訓練データ$D$を浅いNN$g$にバックプロパゲーション学習（backpropagation; BP）させるとは，次の最適化問題を解く

ことである

$$\text{Minimize} \quad \frac{1}{n}\sum_{i=1}^{n} |y_i - g(\boldsymbol{x}_i; \{(\boldsymbol{a}_j, b_j, c_j)\}_{j=1}^{p})|^2 \quad \text{w.r.t.} \quad \{(\boldsymbol{a}_j, b_j, c_j)\}_{j=1}^{p}. \tag{2}$$

2.2. 深層ニューラルネット

まず，ベクトル値の浅い NN を $\boldsymbol{g} : \mathbb{R}^m \to \mathbb{R}^m$ と書く。$\boldsymbol{g}$ の第 k 成分関数 g_k は以下で与えられる

$$g_k(\boldsymbol{x}) := \sum_{j=1}^{p} c_{jk}\sigma(\boldsymbol{a}_j \cdot \boldsymbol{x} - b_j), \quad \boldsymbol{x} \in \mathbb{R}^m. \tag{3}$$

隠れ層パラメータ $(\boldsymbol{a}_j, b_j)$ は k に依らず共通であることに注意せよ。

そして，L 個のベクトル値 NN $\boldsymbol{g}^l : \mathbb{R}^m \to \mathbb{R}^m$ の合成写像を**深層** *NN* という

$$\boldsymbol{g}^{L:l}(\boldsymbol{x}) := \boldsymbol{g}^L \circ \cdots \circ \boldsymbol{g}^1(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathbb{R}^m. \tag{4}$$

本稿で単に深層 NN といえば，$\boldsymbol{g}^{L:l}$（あるいは図 2）のように浅い NN を合成した形式を想定する。

3. 積分表現理論

次の積分で表される関数を NN の積分表現という

$$S[\gamma](\boldsymbol{x}) := \int_{\mathbb{R}^m \times \mathbb{R}} \gamma(\boldsymbol{a}, b)\sigma(\boldsymbol{a} \cdot \boldsymbol{x} - b)\mathrm{d}\lambda(\boldsymbol{a}, b), \quad \boldsymbol{x} \in \mathbb{R}^m. \tag{5}$$

ただし λ は $\mathbb{R}^m \times \mathbb{R}$ 上の Borel 測度とし，λ に対して定まる Hilbert 空間を $\mathcal{G} := L^2(\lambda)$ と書いて，$\gamma \in \mathcal{G}$ とする。典型的には，λ として Lebesgue 測度を想定する。S はパラメータ空間 $\mathcal{G}$ 上の線形作用素とみなせる。

積分表現は，$(\boldsymbol{a}, b) \in \mathbb{R}^m \times \mathbb{R}$ で添字付けられた連続無限個の中間層素子 $\sigma(\boldsymbol{a} \cdot \boldsymbol{x} - b)$ を足し合わせた浅い NN とみなせる。連続化に伴い，出力係数 c_j は係数関数 $\gamma(\boldsymbol{a}, b)$ になる。形式的には，特異測度 $\gamma(\boldsymbol{a}, b)\mathrm{d}\lambda(\boldsymbol{a}, b) = \sum_{j=1}^{p} c_j \mathrm{d}\delta_{\boldsymbol{a}_j, b_j}(\boldsymbol{a}, b)$ をとることで有限の NN も包括的に表現できる。

図 3: 通常の浅い NN（左）と積分表現 NN（右）。積分表現では中間層素子が連続無限個足されている。

NN を積分表現にするメリットはいくつもあるが，最も重要なことは，パラメータが線形になることである。元の NN では非線形関数 σ の中にある隠れ層パラメータ $(\boldsymbol{a}_j, b_j)$ が非線形パラメータとなっていたが，積分表現では隠れ層パラメータが積分消去されるためである。

3.1. バックプロパゲーションの大域最適解

NN の近似対象 $f : \mathbb{R}^m \to \mathbb{R}$ を一つ固定する。f は適当な Hilbert 空間 $\mathcal{F}$ に属しているものとする。例えば，データ分布 $\mu(\boldsymbol{x})$ で重み付けられた Hilbert 空間 $L^2(\mu)$ を想定する。積分表現において，BP 学習は次のように書ける

$$\text{Minimize} \quad \|f - S[\gamma]\|_{\mathcal{F}}^2 + \beta\|\gamma\|_{\mathcal{G}}^2, \quad \text{w.r.t.} \quad \gamma \in \mathcal{G}. \tag{6}$$

ただし $\beta > 0$ とし，議論を簡単にするために L^2 正則化項を設けた。

この積分表現 BP は関数空間上の強凸最適化問題である。形式的に有限次元の場合と同様に計算すると，大域最適解 $\gamma^\star$ が得られる

$$\gamma^\star = (\beta + S^*S)^{-1}S^*f. \tag{7}$$

ただし $S^* : \mathcal{F} \to \mathcal{G}$ は $S : \mathcal{G} \to \mathcal{F}$ の共役作用素である。実は，$S : \mathcal{G} \to \mathcal{F}$ が稠密に定義された閉作用素であれば，この計算が正当化される。通常の BP は局所解との戦いであるから，大域最適解などは夢のようであるが，このようなことが起きるのは γ が線形パラメータだからである。

実際には，S が閉ないし有界となるような $\mathcal{G}, \mathcal{F}$ を定めるのは見た目ほど容易ではない。(Sonoda et al., 2018) では，$\mathcal{F}$ として特殊な再生核 Hilbert 空間（reproducing kernel Hilbert space; RKHS） $H^{\sigma\sigma}$ をとることで，$S : \mathcal{G} \to H^{\sigma\sigma}$ が有界作用素になることを示した。さらに，$H^{\sigma\sigma}$ 上のリッジレット変換 R を用いて

$$\gamma^\star = (\beta + S^*S)^{-1}S^*f = (\beta + 1)^{-1}R[f], \tag{8}$$

と計算できることまで示した。

ここで「$H^{\sigma\sigma}$ 上のリッジレット変換」の定義は準備が長くなるため省略するが，次節で述べる「$L^2(\mathbb{R}^m)$ 上のリッジレット変換」は数値積分を用いて計算可能である。つまり $H^{\sigma\sigma}$ と $L^2(\mathbb{R}^m)$ の厳密な違いに目をつぶれば，$\gamma^\star = (\beta+1)^{-1}R[f]$ を数値積分することで，大域最適な NN $S[\gamma^\star](\boldsymbol{x})$ が得られるということである。

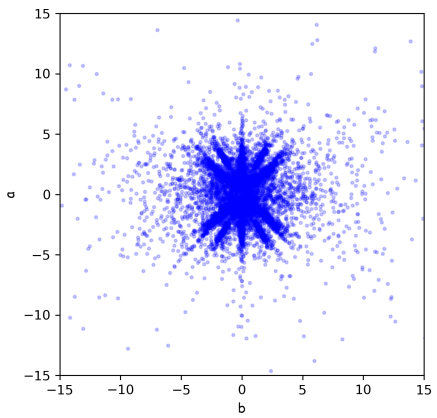

図 4: 学習済 NN の隠れ層パラメータ (a_j, b_j) の散布図

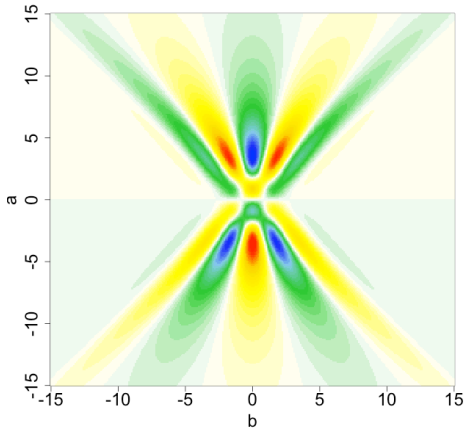

図 5: リッジレット変換 $R[f](a, b)$ のスペクトル

数値実験により，この主張を視覚的に確かめられる。図 4 および図 5 は，同じデータセット $D = \{(x_i, y_i) \mid y_i = f(x_i) = \sin(2\pi x),\ x_i \in \{-1.00, -0.99, \ldots, 0.99, 1.00\}\}$ に対

して行われた2つの実験の結果である。図4は，Dを用いて1,000体の浅いNNをBP学習し，得られた隠れ層パラメータ(a_j, b_j)をプロットした散布図である。乱数初期化された学習前のパラメータは(a, b)空間上にランダムに分布していたはずだが，学習後のパラメータは分布が偏り，全体として星状に分布していることが分かる。つまり，(a, b)空間にはよく使われる場所とそうでない場所があるということである。一方，図5は，数値積分を用いてDから$f(x) = \sin(2\pi x)$のリッジレット変換$R[f](a, b)$を数値計算した結果である。スペクトルは赤が最も高く，黄，緑，青の順に低くなり，緑と青は負の値を示す。リッジレット変換もまた星状の形になることが分かる。そして，2つの図を見較べると，概ねスペクトルの絶対値が大きい場所に，学習済パラメータが分布していることが分かる。BPの大域最適解がリッジレット変換によって得られることを思い出すと，このようなパターンの一致は必ずしも偶然ではないことが理解できる。この現象は筆者が2015年頃に偶然発見したものだが，数学的な証明が付けられたのは2018年になってからである。

3.2. リッジレット変換

以降では，パラメータ空間の測度λをLebesgue測度$\mathrm{d}\lambda(\boldsymbol{a}, b) = \mathrm{d}\boldsymbol{a}\mathrm{d}b$とし，$\mathcal{G} = L^2(\mathbb{R}^m \times \mathbb{R})$および$\mathcal{F} = L^2(\mathbb{R}^m)$とする。

近似対象$f \in \mathcal{F}$を固定し，次の積分方程式を考える

$$\text{Find} \quad \gamma \in \mathcal{G} \quad \text{s.t.} \quad S[\gamma] = f. \tag{9}$$

この方程式はNNの学習問題とみることもできる。通常のNNの場合に，同様の方程式を解くのは至難の技である。ところが積分方程式の場合には，リッジレット変換$\gamma_f = R[f]$が特殊解を与えることが知られている (Murata, 1996; Candès, 1998; Sonoda and Murata, 2017)。即ち，$S[\gamma_f] = S[R[f]] = f$が成り立つ。

ここで，関数$f \in \mathcal{F}$のリッジレット変換は以下で定義される

$$R[f](\boldsymbol{a}, b) := \int_{\mathbb{R}^m} f(\boldsymbol{x})\overline{\rho(\boldsymbol{a} \cdot \boldsymbol{x} - b)}\mathrm{d}\boldsymbol{x}. \tag{10}$$

ただし$\rho : \mathbb{R} \to \mathbb{R}$は急減少関数（$\mathcal{S}$）とし，NNの活性化関数$\sigma \in \mathcal{S}'$に対して次の許容条件（admissibility condition）を満たすものとする

$$\int_{\mathbb{R}\setminus\{0\}} \frac{\widehat{\sigma}(\zeta)\overline{\widehat{\rho}(\zeta)}}{|\zeta|^m}\mathrm{d}\zeta = 1. \tag{11}$$

ここで$\widehat{\cdot}$はFourier変換を表す。許容条件が成り立つ(σ, ρ)の組み合わせにおいて，$R[f]$は$S[\gamma] = f$の特殊解を与える。すなわち，再構成公式が成り立つ

$$S[R[f]] = f, \quad f \in \mathcal{F}. \tag{12}$$

証明は$S[R[f]]$を直接変形することで，Fourier反転公式に帰着することを示す: $S[R[f]] = F^{-1}[\widehat{f}] = f$.

一般に，一つのσに対して許容条件を満たすρは無数に存在する。例えば，σがReLUの場合にはGaussian$\phi(z)$を用いて$\widehat{\rho}(\zeta) = |\zeta|^{m+2+2k}\phi(\zeta)$ $(k = 0, 1, \ldots)$にとれば，ρはいずれも許容条件を満たす (Sonoda and Murata, 2017, § 6)。

4. 輸送理論

次の常微分方程式 (ODE) が定める連続力学系 $\boldsymbol{g}_{t\to s}:\mathbb{R}^m\to\mathbb{R}^m$ を NN のフロー表現という

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{x}_t=\boldsymbol{v}_t(\boldsymbol{x}_t),\quad t\in[0,1]. \tag{13}$$

ただし $\boldsymbol{v}_t(\boldsymbol{x})$ は $(\boldsymbol{x},t)$ における速度ベクトル場を表す。フロー表現にすることで，深層 NN の中間写像の解析を輸送軌道の解析に転嫁できる。

例えばクラス判別をする深層 NN は，最終層が線形判別機になっているので，最終層までにデータ点を線形分離可能な配置にしておかなければならない。従って，中間写像に求められるのは，線形分離できるようにデータ点を輸送することである。このように，深層 NN の機能は輸送軌道の性質として解析するのが自然なケースがある。

フロー表現は，$t\in[0,1]$ で添字付けられた無限個の輸送写像 $\boldsymbol{g}_{t\to t+0}$ を合成して得られる深層 NN とみなせる。Euler 法に代表される折れ線近似を用いて $\boldsymbol{g}^l(\boldsymbol{x})=\boldsymbol{x}+\Delta\boldsymbol{x}^l(\boldsymbol{x})$ に分解することで，有限層の深層 NN が得られる。

図 6: フロー表現 NN の模式図。連続無限個の輸送写像を合成している。

4.0.1. スキップ接続

スキップ接続は，residual network (ResNet) (He et al., 2016) をはじめとする 1,000 層規模の深層 NN を学習するためになくてはならないとされるネットワーク構造である。スキップ接続を式で書くと $\boldsymbol{g}(\boldsymbol{x})=\boldsymbol{x}+\boldsymbol{f}(\boldsymbol{x})$ であり，これは輸送写像である。2018 年には，ResNet を ODE として定式化し，解析ないし一般化する研究が矢継ぎ早に出版された。例えば Lu et al. (2018) は多くの ResNet の亜種が全て，ODE を近似する手法の違いに対応することを指摘した。また Nitanda and Suzuki (2018) は ODE が元の最適化問題を解く過程に対応していることを指摘した。

4.1. 最適輸送写像

フローの特別なクラスとして，最適輸送写像がある。μ_0,μ_1 は $\mathbb{R}^m$ 上の確率測度とし，Lebesgue 測度に対して絶対連続かつ二次モーメントが存在するとする。μ_0 を μ_1 に変形する輸送写像 $\boldsymbol{g}$ の全体を $\mathcal{T}(\mu_0,\mu_1)$ と書く。このとき，次の Monge の最適輸送問題

$$\text{Minimize}\quad \int_{\mathbb{R}^m}|\boldsymbol{x}-\boldsymbol{g}(\boldsymbol{x})|^2\mathrm{d}\mu_0(\boldsymbol{x}),\quad \text{s.t.}\quad \boldsymbol{g}\in\mathcal{T}(\mu_0,\mu_1). \tag{14}$$

の解が一意に存在し，$\mathbb{R}^m$ 上 μ_0-a.e. で一意に定義された強凸関数 $\varphi:\mathbb{R}^m\to\mathbb{R}$ を用いて

$$\boldsymbol{g}_{0\to t}(\boldsymbol{x})=\boldsymbol{x}+t\nabla\varphi(\boldsymbol{x}),\quad t\in[0,1],\ \boldsymbol{x}\in\mathbb{R}^m, \tag{15}$$

と表される（Brenier の極分解定理）。ここで，$\boldsymbol{g}_{0\to t}$ は μ_0 から $\mu_t := (\boldsymbol{g}_{0\to t})_\sharp \mu_0$ への最適輸送写像となっている。また，$t \mapsto \mu_t$ は μ_0 と μ_1 を結ぶ W_2 測地線である。

4.1.1. 生成モデル

Generative adversarial net (GAN) (Goodfellow et al., 2014) は，正規乱数 $\boldsymbol{z}$ を入力し，出力 $\boldsymbol{x} = \boldsymbol{g}(\boldsymbol{z})$ の分布がデータ分布に近づくように生成器 $\boldsymbol{g}$ を訓練する学習法である。このとき $\boldsymbol{g}$ は μ_0 を正規分布とし，μ_1 をデータ分布とする輸送写像である。

GAN に限らず，Variational Autoencoder (VAE) (Kingma and Welling, 2014) や Minimum Probability Flow (Sohl-Dickstein et al., 2011) など多くの生成モデルは輸送写像として解釈可能である。

4.2. 連続方程式と Wasserstein 勾配流

データ分布 μ_0 に従ってデータ点 $\boldsymbol{x}_0$ が分布しているとき，深層 NN 内部の中間データ表現 $\boldsymbol{x}_t = \boldsymbol{g}_{0\to t}(\boldsymbol{x}_0)$ が従う分布 μ_t は次の連続方程式に従って発展する

$$\partial_t \mu_t = -\mathsf{div}\,[\mu_t \boldsymbol{v}_t]. \tag{16}$$

$\mathbb{R}^m$ 上の L^2-Wasserstein 空間を W_2 と書く。速度場 $\boldsymbol{v}_t$ がポテンシャル関数 $V_t : \mathbb{R}^m \to \mathbb{R}$ の勾配として与えられるとき，

$$\boldsymbol{v}_t = \nabla V_t, \quad t \in [0,1] \tag{17}$$

連続方程式 (16) は W_2 上の汎関数 $\mathcal{H}$ による Wasserstein 勾配流に対応付けられる (Villani, 2009, Ex.15.10)

$$\frac{\mathrm{d}}{\mathrm{d}t}\mu_t = -\mathsf{grad}\,\mathcal{H}[\mu_t]. \tag{18}$$

ここで，$\mathcal{H}$ と V_t とは，以下の関係式を満たすように選ぶ

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{H}[\mu_t] = \int_{\mathbb{R}^m} V_t(\boldsymbol{x})\partial_t \mu_t(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \quad \mu_t \in W_2 \tag{19}$$

速度ベクトル場 $\boldsymbol{v}_t$ やフロー表現 $\boldsymbol{g}_t$ は時間に依存するが，Wasserstein 勾配流を特徴付ける汎関数 $\mathcal{H}$ は時間に依存しないので，複雑な輸送過程を扱う場合にも威力を発揮する。

4.2.1. Denoising Autoencoder (DAE)

DAE (Vincent et al., 2008) は，浅い NN $\boldsymbol{g}$ の訓練時にわざとノイズを付加した入力 $\widetilde{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{\varepsilon}$ を与え，ノイズを付加する前の入力 $\boldsymbol{x}$ を推定させる学習法である。DAE の学習則は次のようになる

$$\text{Minimize} \quad \mathbb{E}_{\boldsymbol{x}}\mathbb{E}_{\boldsymbol{\varepsilon}}|\boldsymbol{g}(\boldsymbol{x}+\boldsymbol{\varepsilon}) - \boldsymbol{x}|^2 \quad \text{w.r.t} \quad \boldsymbol{g}. \tag{20}$$

正規雑音 DAE の場合，大域最適解 $\boldsymbol{g}^*$ が陽に求まる (Sonoda and Murata, 2016)

$$\boldsymbol{g}_t^*(\boldsymbol{x}) = \boldsymbol{x} + t\nabla \log[e^{t\triangle}\mu_0](\boldsymbol{x}), \quad t > 0, \boldsymbol{x} \in \mathbb{R}^m. \tag{21}$$

ただし μ_0 は未知のデータ分布とし，t は正規雑音の分散である。証明は変分法を用いて直接計算によって示せるほか，統計的考察からも導ける。式の形から，正規雑音 DAE $\boldsymbol{g}_t^*$ は輸送写像とみなせる。

ここで，$\boldsymbol{v}_0 = \lim_{t\to+0}(\boldsymbol{g}_t^*(\boldsymbol{x}) - \boldsymbol{x})/t = \nabla \log \mu_0(\boldsymbol{x})$ から，極限 $t \to 0$ で正規雑音 DAE の速度ベクトル場は Fisher スコアになることが分かる

$$\lim_{t\to+0} \boldsymbol{v}_t = \nabla \log \mu_0. \tag{22}$$

従って，$\boldsymbol{g}_t^*$ によるデータ分布 μ_0 の押出分布を μ_t と書くことにすると，連続方程式から μ_t は逆拡散方程式に従って発展することが分かる

$$\partial_t \mu_t|_{t=0} = -\mathsf{div}\,[\mu_0 \boldsymbol{v}_0] = -\triangle \mu_0. \tag{23}$$

そして，熱方程式に対応する Wasserstein 勾配流は Shannon エントロピー $\mathcal{H}[\mu] := -\int_{\mathbb{R}^m} \mu(\boldsymbol{x}) \log \mu(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$ を汎関数とする勾配流であることを使うと，さらに

$$\frac{\mathrm{d}}{\mathrm{d}t}\mu_t\Big|_{t=0} = -\mathsf{grad}\,\mathcal{H}[\mu_0]. \tag{24}$$

となることが分かる。即ち，μ_t は Wasserstein 空間上でエントロピーを減らす方向に流れている。

ここまでの議論は $t \to 0$ に限るが，DAE から得られたデータ分布に対して繰り返し DAE を適用する方法で得られる合成 DAE において，各 DAE の分散 $t \to 0$ の極限をとると，深層正規雑音 DAE の連続極限である連続 DAE $\boldsymbol{g}_{0\to t}\,(t > 0)$ が得られる。連続 DAE は，各時刻 t において $\boldsymbol{v}_t = \nabla \log \mu_t$，$\partial_t \mu_t = -\triangle \mu_t$，$\frac{\mathrm{d}}{\mathrm{d}t}\mu_t = -\mathcal{H}[\mu]$ を満たすことが示せる。すなわち，正規雑音 DAE のフロー表現はデータ分布のエントロピーを減らすようにデータ点を輸送する深層 NN である。

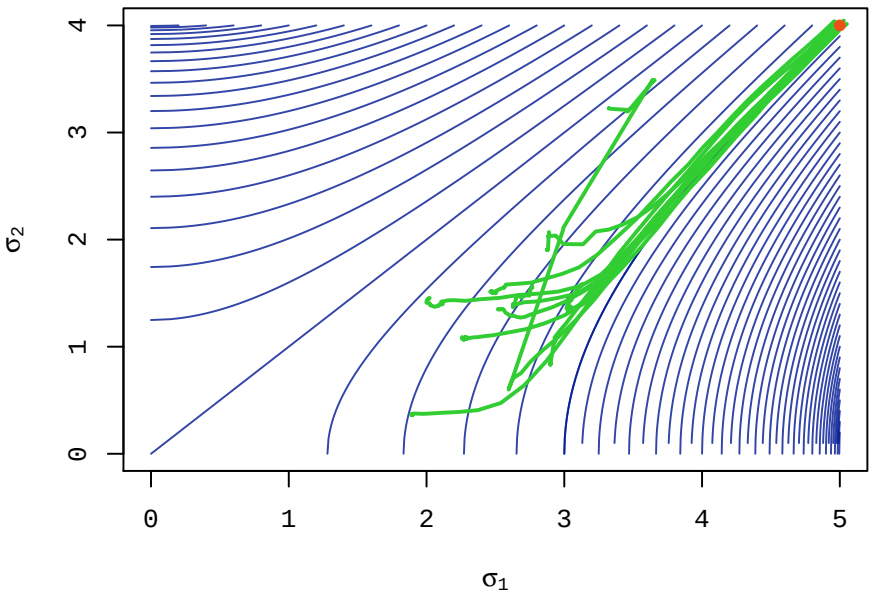

図 7: 2 次元正規分布の部分空間 $\mathcal{Q}$ 上のエントロピー勾配流（青）と，実際に学習された合成 DAE による軌跡（緑）

図 7 は，確率分布の空間 $\mathcal{Q}$ 上に，理論的に計算されたエントロピー勾配流と，実際に合成 DAE を学習して得られた軌跡をプロットしたものである。特に右上の方では理

論と実験が一致している。ここで $\mathcal{Q}$ は，平均 $(0,0)$ 共分散行列が $V_{11}=\sigma_1^2, V_{12}=V_{21}=0, V_{22}=\sigma_2^2$ で与えられる 2 次元正規分布の空間である。(σ_1,σ_2) 座標系において W_2 距離はユークリッド距離と一致するため，勾配流は曲率を含めて可視化できる。

5. まとめ

積分表現理論と輸送理論から NN のホワイトボックス化を試みた。積分表現理論によれば，大域最適な浅い NN のパラメータは，データのリッジレット変換で与えられる。輸送理論によれば，深層 NN の中間写像は ODE を用いて表現できる。フローが求まれば，各中間写像はリッジレット変換を通じて浅い NN に翻訳可能である。ホワイトボックス化という観点では，機能が同じであれば実装は NN でなくとも良い。輸送理論の方は，大域最適を与えるリッジレット変換のような道具立てがまだ揃っていないので，今後の発展が期待される。

参考文献

E. J. Candès. *Ridgelets: theory and applications*. PhD thesis, Standford University, 1998.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680, Montreal, BC, 2014.

K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, Nevada, 2016. IEEE.

G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury. Deep Neural Networks for Acoustic Modeling in Speech Recognition. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations 2014*, pages 1–14, Banff, BC, 2014.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105, Lake Tahoe, 2012.

Y. Lu, A. Zhong, Q. Li, and B. Dong. Beyond Finite Layer Neural Networks: Bridging Deep Architectures and Numerical Differential Equations. In *Proceedings of The 35th International Conference on Machine Learning*, number 80, 2018.

C. Molnar. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. URL https://christophm.github.io/interpretable-ml-book/index.html.

N. Murata. An integral representation of functions using three-layered betworks and their approximation bounds. *Neural Networks*, 9(6):947–956, 1996.

A. Nitanda and T. Suzuki. Functional Gradient Boosting based on Residual Network Perception. In *Proceedings of The 35th International Conference on Machine Learning*, 2018.

A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations 2016*, pages 1–15, San Juan, Puerto Rico, 2016.

J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You Only Look Once: Unified, Real-Time Object Detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.

M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter. Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 1528–1540, 2016.

D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations 2015*, pages 1–14, San Diego, California, 2015.

J. Sohl-Dickstein, P. Battaglino, and M. R. DeWeese. Minimum Probability Flow Learning. In *Proceedings of The 28th International Conference on Machine Learning (ICML-11)*, pages 905–912, Belleview, WA, USA, 2011. ACM.

S. Sonoda and N. Murata. Decoding Stacked Denoising Autoencoders. 2016.

S. Sonoda and N. Murata. Neural network with unbounded activation functions is universal approximator. *Applied and Computational Harmonic Analysis*, 43(2):233–268, 2017.

S. Sonoda, I. Ishikawa, M. Ikeda, K. Hagihara, Y. Sawano, T. Matsubara, and N. Murata. Integral representation of the global minimizer. 2018.

C. Szegedy, W. Zaremba, and I. Sutskever. Intriguing properties of neural networks. In *International Conference on Learning Representations 2014*, pages 1–10, Banff, BC, 2014.

A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. WaveNet: A Generative Model for Raw Audio. 2016.

C. Villani. *Optimal Transport: Old and New*. Springer-Verlag, 2009.

P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and Composing Robust Features with Denoising Autoencoders. In *Proceedings of The 25th International Conference on Machine Learning (ICML-08)*, pages 1096–1103, Helsinki, 2008. ACM.

O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2015.

Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Ł. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean. Google's neural machine translation system: bridging the gap between human and machine translation. 2016.

L. Xiao, Y. Bahri, J. S.-d. Samuel, and S. S. Jeffrey. Dynamical Isometry and a Mean Field Theory of CNNs: How to Train 10,000-Layer Vanilla Convolutional Neural Networks. In *Proceedings of The 35th International Conference on Machine Learning*, volume 80, pages 389–5398, Stockholm, 2018. PMLR.