深層学習の Wasserstein 幾何学的解析にむけた取組み *

園田　翔 [†]　村田　昇
早稲田大学　先進理工学部
Sonoda Sho　Noboru Murata
School of Advanced Science and Engineering
Waseda University

1　はじめに

深層ニューラルネットはどのように情報を処理しているのだろうか？ 発表者らが推進する**深層ニューラルネットの輸送解析**では，ニューラルネットの中間層を輸送写像とみなし，輸送の性質によってニューラルネットを分類する。別の言い方をすれば，多層の中間層を特徴量写像の合成写像が為す力学系として捉え，輸送軌道によってニューラルネットを特徴付ける。ニューラルネットのパラメータは値が異なっていても同じ写像を表すことがあるので，パラメータに基づくニューラルネットの解析は困難である。また，パラメータに沿った最小二乗法では大域最適を見つけることも困難である。一方，輸送軌道はパラメータとは独立な幾何学的対象であるので，ニューラルネットの写像としての性質を調べるのに有効である。

これまでに，Gaussian denoising autoencoder (DAE) と呼ばれる深層ニューラルネットの一種が，データ分布のエントロピーを減らす方向に輸送する写像であることが分かっている。本研究の目的は，一般の雑音分布による DAE や，教師あり学習による深層ニューラルネットを輸送写像として記述することである。

2　Denoising Autoencoder

Denoising Autoencoder （DAE）は，深層ニューラルネットの輸送解析を動機付ける基本的なクラスである。DAE とは，訓練データにわざと雑音を加え，雑音を除去するようにニューラルネットを訓練するオートエンコーダーの亜種である [Vincent et al., 2008]。

---

* 科研費シンポジウム「大規模複雑データの理論と方法論，及び，関連分野への応用」

† sho.sonoda@aoni.waseda.jp

2.1 DAE の学習手続き

$\boldsymbol{x} \sim \pi$ を $\mathbb{R}^m$ に値をとるデータ，$\boldsymbol{\varepsilon} \sim \nu_t$ を平均 0 分散共分散行列 tI の加法雑音として，$\widetilde{\boldsymbol{x}}$ を雑音が付加された観測データ

$$\widetilde{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{\varepsilon} \tag{1}$$

とする。DAE の学習手続きでは，学習機 $\boldsymbol{g}$ に $\widetilde{\boldsymbol{x}}$ を提示して，$\boldsymbol{x}$ を推定させる。

本来，雑音は学習機のロバスト性を強化したり，データを水増ししたりする目的で導入されたが，このように雑音を除去する学習法によって訓練された学習機は，結果として雑音を補正する機能を獲得することになる。本研究では，この補正項を輸送作用とみなして，深層ニューラルネットの特徴付けを行う。

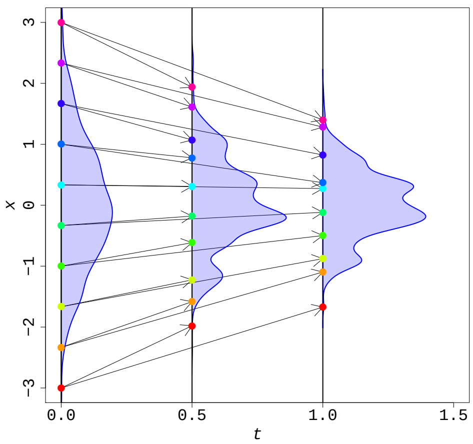

図 1　DAE はデータ点を輸送する写像とみなせる。左から元のデータ分布 π_0，$\boldsymbol{g}_{0.5}$ によって輸送されたデータ点の分布（押出測度）$\pi_{0.5}$，$\boldsymbol{g}_{1.0}$ によって輸送されたデータ点の分布 $\pi_{1.0}$。t 軸は雑音の分散を表し，輸送解釈では輸送時間に対応する。x 軸はデータ空間 $\mathbb{R}^1$ を表す。データ分布は正規分布 $\mathcal{N}(0,1)$，雑音分布も正規分布 $\mathcal{N}(0,t)$ である。雑音が強いほど，押出測度の分散は小さくなる。本研究の解析を通じて，正規雑音の場合には，厳密には分散ではなくエントロピーを減らすように輸送していることが明らかとなる。

2.2 DAE の輸送写像（変分法による定式化）

通常，ニューラルネットは最小二乗法によって学習させるので，DAE は次の最適化問題（変分問題）と等価である:

$$\min_{\boldsymbol{g}} \mathbb{E}_{\pi} \mathbb{E}_{\nu_t} |\boldsymbol{g}(\boldsymbol{x} + \boldsymbol{\varepsilon}) - \boldsymbol{x}|^2. \tag{2}$$

ただし $\boldsymbol{g}$ は十分広いクラスの関数を表現でき，停留点を達成できるものとする。

この変分問題は変分計算を用いて停留点 $\boldsymbol{g}_t^*$ を求めることができ，以下のようになる [Sonoda and Murata, 2016, 2017]

$$\boldsymbol{g}_t^*(\boldsymbol{x}) = \boldsymbol{x} - \frac{1}{\nu_t * \pi(\boldsymbol{x})} \int_{\mathbb{R}^m} \boldsymbol{\varepsilon} \nu_t(\boldsymbol{\varepsilon}) \pi(\boldsymbol{x} - \boldsymbol{\varepsilon}) \mathrm{d}\boldsymbol{\varepsilon}. \tag{3}$$

ただし $*$ は畳み込み積分を表す:

$$\nu_t * \pi(\boldsymbol{x}) = \int_{\mathbb{R}^m} \nu_t(\boldsymbol{\varepsilon}) \pi(\boldsymbol{x} - \boldsymbol{\varepsilon}) \mathrm{d}\boldsymbol{\varepsilon}. \tag{4}$$

DAE の式 (3) の意味を考えてみよう。まず，第一項は恒等写像であるから，オートエンコーダーとしての性質を表している。一方，第二項は雑音除去に伴って現れた補正項である。従って，DAE は**データ点 $\boldsymbol{x}$ を補正項の方向に輸送する輸送写像**とみなせる。

以下では (3) の導出を説明する。まず，変分問題の目的関数を汎関数 $L[\boldsymbol{g}] := \mathbb{E}_\pi \mathbb{E}_{\nu_t} |\boldsymbol{g}(\boldsymbol{x} + \boldsymbol{\varepsilon}) - \boldsymbol{x}|^2$ とおく。次に，任意の関数 $\boldsymbol{h}$ に対し，$\boldsymbol{g}$ における汎関数 L の $\boldsymbol{h}$ 方向変分 $\delta L_{\boldsymbol{h}}[\boldsymbol{g}]$ を計算する。最初に，L の積分変数を変換し，積分順序を変更する

$$\begin{aligned} L[\boldsymbol{g}] &= \int_{\mathbb{R}^m} \int_{\mathbb{R}^m} |\boldsymbol{g}(\boldsymbol{x} + \boldsymbol{\varepsilon}) - \boldsymbol{x}|^2 \nu_t(\boldsymbol{\varepsilon}) \pi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{\varepsilon} \\ &= \int_{\mathbb{R}^m} \int_{\mathbb{R}^m} |\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{x} + \varepsilon|^2 \nu_t(\boldsymbol{\varepsilon}) \pi(\boldsymbol{x} - \boldsymbol{\varepsilon}) \mathrm{d}\boldsymbol{\varepsilon} \mathrm{d}\boldsymbol{x}. \end{aligned}$$

すると，$\delta L_{\boldsymbol{h}}[\boldsymbol{g}]$ は以下のように計算される

$$\begin{aligned} \delta L_{\boldsymbol{h}}[\boldsymbol{g}] &= \frac{\mathrm{d}}{\mathrm{d}s} L[\boldsymbol{g} + s\boldsymbol{h}] \Big|_{s=0} \\ &= 2 \int_{\mathbb{R}^m} \left[\int_{\mathbb{R}^m} [\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{x} + \varepsilon] \nu_t(\boldsymbol{\varepsilon}) \pi(\boldsymbol{x} - \boldsymbol{\varepsilon}) \mathrm{d}\varepsilon \right] \boldsymbol{h}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \end{aligned}$$

停留点 $\boldsymbol{g}_t^*$ においては任意の $\boldsymbol{h}$ 方向に対して $\delta L_{\boldsymbol{h}}[\boldsymbol{g}_t^*] \equiv 0$ が成り立つ。従って，変分法の基本補題によって被積分関数がほとんど至る所 0 になることが言える

$$\int_{\mathbb{R}^m} [\boldsymbol{g}_t^*(\boldsymbol{x}) - \boldsymbol{x} + \varepsilon] \nu_t(\boldsymbol{\varepsilon}) \pi(\boldsymbol{x} - \boldsymbol{\varepsilon}) \mathrm{d}\varepsilon = 0, \quad \text{a.e. } \boldsymbol{x} \in \mathbb{R}^m. \tag{5}$$

この方程式を $\boldsymbol{g}_t^*$ に関して解いて，変分問題 (2) の停留点として (3) が得られる

2.3 DAE の輸送写像（統計的推定問題として定式化）

DAE の学習手続きは**平均の推定**と等価であることに注意すると，DAE 輸送写像 (3) は事後平均（posterior mean）$\mathbb{E}[\boldsymbol{x} \mid \widetilde{\boldsymbol{x}}]$ であることが分かる。すなわち，事後平均は

$$\begin{aligned} \mathbb{E}[\boldsymbol{x} \mid \widetilde{\boldsymbol{x}}] &= \frac{\int_{\mathbb{R}^m} \boldsymbol{x} p(\widetilde{\boldsymbol{x}} \mid \boldsymbol{x}) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}}{\int_{\mathbb{R}^m} p(\widetilde{\boldsymbol{x}} \mid \boldsymbol{x}') p(\boldsymbol{x}') \mathrm{d}\boldsymbol{x}'} \\ &= \frac{\int_{\mathbb{R}^m} \boldsymbol{x} \nu_t(\widetilde{\boldsymbol{x}} - \boldsymbol{x}) \pi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}}{\int_{\mathbb{R}^m} \nu_t(\widetilde{\boldsymbol{x}} - \boldsymbol{x}') \pi(\boldsymbol{x}') \mathrm{d}\boldsymbol{x}'} \end{aligned} \tag{6}$$

$$
\begin{aligned}
&= \frac{1}{\nu_t * \pi(\widetilde{\boldsymbol{x}})} \int_{\mathbb{R}^m} (\widetilde{\boldsymbol{x}} - \boldsymbol{\varepsilon}) \nu_t(\boldsymbol{\varepsilon}) \pi(\widetilde{\boldsymbol{x}} - \boldsymbol{\varepsilon}) \mathrm{d}\boldsymbol{\varepsilon}, \quad \boldsymbol{x} \leftarrow \widetilde{\boldsymbol{x}} - \boldsymbol{\varepsilon} \\
&= \widetilde{\boldsymbol{x}} - \frac{1}{\nu_t * \pi(\widetilde{\boldsymbol{x}})} \int_{\mathbb{R}^m} \boldsymbol{\varepsilon} \nu_t(\boldsymbol{\varepsilon}) \pi(\widetilde{\boldsymbol{x}} - \boldsymbol{\varepsilon}) \mathrm{d}\boldsymbol{\varepsilon},
\end{aligned} \tag{7}
$$

となるので，$\boldsymbol{g}_t^*(\widetilde{\boldsymbol{x}}) = \mathbb{E}[\boldsymbol{x} \mid \widetilde{\boldsymbol{x}}]$ が分かる。なお，事後平均の途中式として現れた (6) は，雑音が加法的でない場合にも成り立つ式であり，Alain and Bengio [2014] はこの形式を用いて DAE を解析した。一方， (3) は第一項（$\boldsymbol{x}$）が出発地点，第二項が変位ベクトルに対応し，全体として出発地点から変位ベクトルの分だけ移動する，という輸送写像としての形式が陽に現れている。本研究では，後者の形式を軸として解析を行う。縮小推定量の文脈などでは，(3) は 事後平均の Brown 表現（Brown's representation of the posterior mean）としても知られている [George et al., 2006]。

2.4　正規雑音 DAE の輸送写像

正規雑音の場合には (3) をさらに簡約化できる [Sonoda and Murata, 2016]。まず，雑音分布を平均 0 分散共分散行列 tI の正規分布とする

$$
\nu_t(\boldsymbol{\varepsilon}) = \frac{1}{(2\pi t)^{m/2}} \exp\left(-\frac{|\boldsymbol{\varepsilon}|^2}{2t}\right). \tag{8}
$$

このとき，以下の恒等式（Stein's identity）が成り立つ

$$
\boldsymbol{\varepsilon}\nu_t(\boldsymbol{\varepsilon}) = -t\nabla\nu_t(\boldsymbol{\varepsilon}). \tag{9}
$$

この恒等式が成り立つのは正規分布の場合に限る。

従って，(3) は以下のように簡略化できる

$$
\begin{aligned}
\boldsymbol{g}_t^*(\boldsymbol{x}) &= \boldsymbol{x} - \frac{1}{\pi * \nu_t(\boldsymbol{x})} \int_{\mathbb{R}^m} \boldsymbol{\varepsilon}\pi(\boldsymbol{x} - \boldsymbol{\varepsilon})\nu_t(\boldsymbol{\varepsilon})\mathrm{d}\boldsymbol{\varepsilon} \\
&= \boldsymbol{x} + \frac{t}{\pi * \nu_t(\boldsymbol{x})} \int_{\mathbb{R}^m} \pi(\boldsymbol{x} - \boldsymbol{\varepsilon})\nabla\nu_t(\boldsymbol{\varepsilon})\mathrm{d}\boldsymbol{\varepsilon} \\
&= \boldsymbol{x} + \frac{t\nabla\pi * \nu_t(\boldsymbol{x})}{\pi * \nu_t(\boldsymbol{x})} \\
&= \boldsymbol{x} + t\nabla\log[\pi * \nu_t](\boldsymbol{x}).
\end{aligned} \tag{10}
$$

この式の意味するところは以下のとおりである。すなわち，正規雑音 DAE において雑音が付加されたデータの分布は $\pi * \nu_t$ で与えられるので，そのスコア $-\nabla\log[\pi * \nu_t]$ に相当する量を補正せよという意味である。

3　輸送に伴うデータ分布の変化

前節の解析を通じて，深層ニューラルネットを輸送写像としてモデル化する動機付けを行った。本節では一旦 DAE を離れ，一般の輸送写像 $\boldsymbol{g}_t$ とデータ分布 π_0 が与えられた場合

に，データ点 $\boldsymbol{x}$ の輸送に伴ってデータ分布が変形していく過程について説明する。変形後のデータ分布は押出測度と呼ばれるものであり，$\boldsymbol{g}_{t\sharp}\pi_0$ と書くこともあるが，誤解の恐れのない限りは単に π_t と書く。

3.1　連続方程式

データ点の輸送に伴い，全データ点の総量は保存するものとする。この条件は，π_0 と π_t がともに確率測度になるために必要である。位置 $\boldsymbol{x}$ 時刻 t における輸送速度を $\boldsymbol{v}_t(\boldsymbol{x})$ と書くことにする。質量が保存するという仮定から直ちに，π_t の時間発展法則は連続方程式（continuity equation）

$$\partial_t \pi_t(\boldsymbol{x}) = -\nabla \cdot [\pi_t(\boldsymbol{x})\boldsymbol{v}_t(\boldsymbol{x})], \quad \boldsymbol{x} \in \mathbb{R}^m,\ t \geq 0 \tag{11}$$

に従うことが分かる[*1]。ここで，$\nabla\cdot$ は発散作用素を表す。

3.2　連続方程式を導く Wasserstein 勾配流

連続方程式と Wasserstein 勾配流との間には対応付けが知られている。従って，輸送写像は連続方程式という偏微分方程式だけでなく，Wasserstein 勾配流という発展方程式としても特徴付けができる。つまり，連続方程式に従ってデータ分布 π_t が時間発展する様子は，Wasserstein 空間上の曲線（軌道）$t \mapsto_t$ として理解できる。Wasserstein 勾配流の基礎づけについては補遺を参照せよ。

連続方程式 (11) において，ベクトル場 $\boldsymbol{v}_t$ 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場合を考える:

$$\boldsymbol{v}_t(\boldsymbol{x}) = \nabla V_t(\boldsymbol{x}), \quad t > 0,\, \boldsymbol{x} \in \mathbb{R}^m. \tag{12}$$

さらに，$\mathcal{F}$ と V_t とは，以下の関係式を満たすように選ばれているものとする

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}[\pi_t] = \int_{\mathbb{R}^m} V_t(\boldsymbol{x})\partial_t \pi_t(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \quad \pi_t \in \mathcal{W}_2(\mathbb{R}^m) \tag{13}$$

このとき，Wasserstein 幾何学の基本的な結果 [Villani, 2009, Ex.15.10] により，$\mathcal{F}$ による Wasserstein 勾配流 π_t は，∇V_t による連続方程式を満たすことが知られている。

$$\partial_t \pi_t(\boldsymbol{x}) = -\nabla \cdot [\pi_t(\boldsymbol{x})\nabla V_t(\boldsymbol{x})], \quad t \geq 0,\, \boldsymbol{x} \in \mathbb{R}^m. \tag{14}$$

連続方程式を特徴付ける速度場 $\boldsymbol{v}_t$ は時間に依存するが，Wasserstein 勾配流を特徴付けるポテンシャル汎関数 $\mathcal{F}$ 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ことが期待される。実際，後に例として述べる通り，ポテンシャル汎関数にはエントロピーなどのよく知られた汎関数が登場する。

[*1] 厳密には π_t は確率密度関数の変数変換の公式を通じて計算する。

4　正規雑音 DAE の輸送解析

正規雑音 DAE を例に取り，輸送解析を行う。

4.1　初速度ベクトルの解析

まず，正規雑音 DAE 輸送写像 (10) の “初速度” ベクトルは次のように計算できる

$$
\begin{aligned}
\partial_t \boldsymbol{g}^*_{t=0}(\boldsymbol{x}) &= \nabla \log[\pi * \nu_{t=0}](\boldsymbol{x}) + 0 \cdot \nabla \log[\pi * \partial_t \nu_{t=0}](\boldsymbol{x}) \\
&= \nabla \log \pi(\boldsymbol{x}). \qquad (15)
\end{aligned}
$$

ただし，正規分布の性質 $\lim_{t\to 0} \nu_t = \delta$ を用いた。すなわち，**正規雑音 *DAE* の場合，輸送に伴う初速度ベクトルはスコア**で与えられるのである。この性質は $t \to 0$ に限ることに注意せよ。つまり，一般の $t > 0$ の場合には初速度ベクトルはスコアにはならない。

次に，$t \to 0$ におけるデータ分布の変形法則が**逆拡散方程式**に従うことを示す。正規雑音 DAE 輸送写像の初速ベクトル場は (15) で与えられたので，連続方程式 (11) に $\boldsymbol{v}_0 = \nabla \log \pi_0$ を代入して以下を得る

$$
\begin{aligned}
\partial_t \pi_{t=0}(\boldsymbol{x}) &= -\nabla \cdot [\pi_0(\boldsymbol{x}) \nabla \log \pi_0(\boldsymbol{x})] \\
&= -\nabla \cdot [\nabla \pi_0(\boldsymbol{x})] \\
&= -\triangle \pi_0(\boldsymbol{x}). \qquad (16)
\end{aligned}
$$

すなわち，**正規雑音 *DAE* の場合，$t \to 0$ における輸送に伴うデータ分布の変形法則は逆拡散方程式 $\partial_t \pi_{t=0}(\boldsymbol{x}) = -\triangle \pi_0(\boldsymbol{x})$** に従うのである。

最後に，正規雑音 DAE はエントロピーを減らすようにデータ点を輸送する写像であることを示す。熱方程式 $\partial_t \pi_t = \triangle \pi_t$ の解はエントロピー汎関数

$$
\mathcal{H}[\pi] := -\int_{\mathbb{R}^m} \pi(\boldsymbol{x}) \log \pi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \qquad (17)
$$

を汎関数とする Wasserstein 勾配流 $\frac{\mathrm{d}}{\mathrm{d}t}\pi_t = \mathsf{grad}\, \mathcal{H}[\pi_t]$ の解であることが知られている。実際，(15) からポテンシャル関数は $V_t = \log \pi_t$ となることが予想され，直ちに (13) を満たすことが分かる。このことから，逆拡散方程式 (16) はエントロピーを減らす勾配流

$$
\frac{\mathrm{d}}{\mathrm{d}t}\pi_0 = -\mathsf{grad}\, \mathcal{H}[\pi_0] \qquad (18)
$$

に対応することが分かる。すなわち，DAE 輸送写像 $\boldsymbol{g}^*$ は**データ分布 π のエントロピーを減らすように輸送する**写像である。

逆拡散やエントロピー減少の意味を考えてみよう。逆拡散方程式において，時間変数の正負を反転すると通常の熱方程式が得られる。従って，逆拡散方程式は拡散現象を時間逆向きに遡る現象を記述している。自然界の多くの系は，エントロピーが増大する方向に発展するので，これは不自然な現象である。このようなことが起こるのは，DAE の訓練過程が雑音除去であり，本質的に逆問題を解いていることから納得される。

4.2　深層正規雑音 DAE

ここまで，$t \to 0$ の極限を考えてきた。一般の時刻 $t > 0$ においては，正規雑音 DAE と言えども，スコアや逆拡散方程式，エントロピー勾配流のような著しい性質は損なわれる。このことを Wasserstein 空間 $\mathcal{W}_2(\mathbb{R}^m)$ 上で考えてみよう。正規雑音 DAE に伴うデータ分布 π_t の時間発展は，初速のみエントロピー勾配方向 $\mathsf{grad}\,\mathcal{H}[\pi_t]$ を向いているが，時刻 t の発展とともに次第に勾配流から乖離していく。

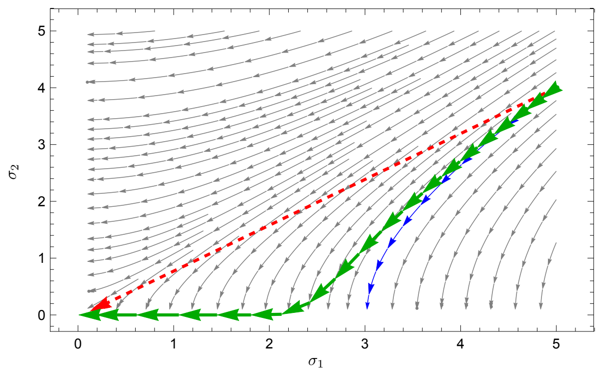

図 2　正規雑音 DAE の時間発展法則はエントロピーを汎関数とする Wasserstein 空間上の勾配流（グレー）になっている。ただし，DAE が勾配流に沿うのは初速のみであり，その後は勾配流から乖離する（赤）。繰り返し DAE の訓練と合成を繰り返すことにより，エントロピー勾配流に近い折れ線が得られる（緑）。これが深層 DAE である。無限小時間 DAE を無限に合成した極限として，各点がエントロピー勾配流に一致している連続 DAE が得られる（青）。

そこで，一回の正規雑音 DAE を短時間 Δt に抑えて，得られたデータ分布に対して再び正規雑音 DAE を適用することを考える。すると，L 回の繰り返しで L 層の深層 DAE

$$\boldsymbol{g}^{L:1} := \boldsymbol{g}^{L} \circ \cdots \circ \boldsymbol{g}^{1} \tag{19}$$

が得られる。特に，各層において初速はエントロピー勾配流に従う。こうして得られる $\mathcal{W}_2(\mathbb{R}^m)$ 上の軌跡は，エントロピー勾配流の折れ線近似（接線近似，Euler 近似）である。

さらに，$\Delta t \to 0$ の無限小極限では，連続無限層の DAE $\boldsymbol{g}_t$ が得られる。すなわち，連続無限層正規雑音 DAE は，任意の時刻 $t \geq 0$ において以下の性質を満たす理想的な輸送写像である

$$\partial_t \boldsymbol{g}_t(\boldsymbol{x}) = \nabla \log \pi_t(\boldsymbol{x}), \tag{20}$$
$$\partial_t \pi_t(\boldsymbol{x}) = -\triangle \pi_t(\boldsymbol{x}), \tag{21}$$
$$\frac{\mathrm{d}}{\mathrm{d}t}\pi_t = -\mathsf{grad}\,\mathcal{H}[\pi_t]. \tag{22}$$

一般に，始点と終点は等しいが，途中の輸送経路が異なるような連続輸送写像は無数に存在する。このような不定性により，深層ニューラルネットは学習過程で輸送経路が迷子にな

る可能性が高い。連続 DAE は，エントロピー勾配流（連続力学系）によって各時刻での振舞いまで規定されているので，輸送経路の不定性が排除されている。従って，エントロピー汎関数は深さ方向の正則化項の役割を担っていることが分かる。

5　一般の深層ニューラルネットの輸送解析に向けて

本節では，一般の深層ニューラルネットを取り上げ，輸送解析の展望を紹介する。

5.1　Tsallis エントロピー勾配流

Tsallis エントロピーは以下で与えられる

$$\mathcal{H}^q[\pi] := -\int_{\mathbb{R}^m} \frac{\pi^q(x) - \pi(x)}{q-1} \mathrm{d}x.$$

このとき，$V_t = -\frac{q}{q-1}\pi_t^{q-1}$ であり，$\mathrm{grad}\,\mathcal{H}^q[\pi_t](x) = \triangle \pi_t^q(x)$ が分かる。従って，$-\mathrm{grad}\,\mathcal{H}^q$ に対応する連続方程式は “逆” 多孔媒質方程式（backward porous medium equation）である [Villani, 2009, Ex.15.6]

$$\partial_t \pi_t = -\triangle \pi_t^q. \tag{23}$$

このとき，雑音分布は q-正規分布になることが予想されるが，証明には至っていない。

5.2　教師有りニューラルネット

教師有り学習では，各データ点 $\boldsymbol{x}$ にラベルが付与されている。輸送に伴い，同じラベル同士の点は近づき，異なるラベル同士の点は遠ざかることが期待される。このような輸送現象は多成分系の拡散現象として記述できる。

$$\partial_t \boldsymbol{\mu}_t = \triangle \boldsymbol{\mu}_t + \mathbf{R}(\boldsymbol{\mu}_t). \tag{24}$$

ただし $\boldsymbol{\mu}_t$ の各成分はラベル毎の質量濃度を表す。

5.3　ConvNet

ResNet や Highway Network に見られるスキップコネクションは，ConvNet を著しく深くするためのヒューリスティクスとして不可欠である。スキップコネクションは輸送写像を陽に表した形式 $\boldsymbol{x} + \mathbf{f}(\boldsymbol{x})$ であるから，ConvNet においても輸送解析が展開できることを示唆している。

6　まとめ

輸送写像とみなすことで，深層ニューラルネットは $\mathbb{R}^m$ 上や Wasserstein 空間 $\mathcal{W}_2(\mathbb{R}^m)$ 上の軌道に対応付けられる。通常の深層ニューラルネットは，連続ニューラルネットの折れ線近似とみなせる。正規雑音 DAE の例では，輸送写像が解析的に求まることを示し，エントロピー汎関数が深さ方向の正則化を担っていることを見た。一般の輸送写像の場合にも，学習の手続きと，輸送を規定する汎関数との対応を付けることで，深層学習の解釈性向上に貢献することが期待される。

付録 A　連続方程式と Wasserstein 勾配流

$\mathbb{R}^m$ 上の L^2-Wasserstein 空間 $\mathcal{W}_2(\mathbb{R}^m)$ とは，$\mathbb{R}^m$ 上の絶対連続かつ 2 次モーメントが存在する確率測度の空間に，L^2-Wasserstein 計量 $\mathfrak{g}$ という無限次元 Riemann 計量を導入した関数多様体である。例えば，一つの確率分布 π は $\mathcal{W}_2(\mathbb{R}^m)$ の一点に相当する。パラメタ付けられた確率分布族（$\mathcal{W}_2(\mathbb{R}^m)$ 上の曲線）π_t $t \in [0,1]$ の時間微分 $\partial_t \pi_{t=0}$ は，π_0 における接ベクトル $\dot{\pi}_0$ に対応する。特に，$\mathcal{W}_2(\mathbb{R}^m)$ は Riemann 多様体なので，勾配作用素 grad や勾配流が定義できる [桑江一洋 et al., 2015]。

Wasserstein 空間 $\mathcal{W}_2(\mathbb{R}^m)$ 上の汎関数 $\mathcal{F}$ による Wasserstein 勾配流とは，以下の発展方程式の解 π_t である

$$\frac{\mathrm{d}}{\mathrm{d}t}\pi_t = \mathsf{grad}\,\mathcal{F}[\pi_t], \quad \pi_t \in \mathcal{W}_2(\mathbb{R}^m). \tag{25}$$

ただし π_t が $\mathcal{W}_2(\mathbb{R}^m)$ の点であることを強調して，偏微分 ∂_t ではなく常微分 $\mathrm{d}/\mathrm{d}t$ を用いている。

参考文献

G. Alain and Y. Bengio. What Regularized Auto-Encoders Learn from the Data Generating Distribution. *Journal of Machine Learning Research*, 15:3743–3773, 2014.

E. I. George, F. Liang, and X. Xu. Improved minimax predictive densities under Kullback–Leibler loss. *Annals of Statistics*, 34(1):78–91, 2006.

S. Sonoda and N. Murata. Decoding Stacked Denoising Autoencoders. arXiv:1605.02832 2016.

S. Sonoda and N. Murata. Transportation analysis of denoising autoencoders: a novel method for analyzing deep neural networks. In *NIPS Workshop on Optimal Transport & Machine Learning*, pages 1–10, Long Beach, 2017.

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

P. Vincent, H. Larochelle, Y. Bengio, and P. -A. Manzagol. Extracting and Composing Robust Features with Denoising Autoencoders. In *ICML-08*, pages 1096–1103, Helsinki, 2008.

桑江一洋, 塩谷隆, 太田慎一, 高津飛鳥, and 桑田和正. 最適輸送理論とリッチ曲率. In 中央大学理工学部数学教室, editor, 第 *63* 回 *ENCOUNTER with MATHEMATICS*, page 115, 2015.