第 17 回数学総合若手研究集会 講演テクニカルレポート

Wasserstein 勾配流に対する差分法と深層学習

東京大学大学院情報理工学系研究科数理情報学専攻修士一年
磯部 伸 (Noboru ISOBE)

概要

深層ニューラルネットワークを力学系的に理解しようとする試みは ODENet [1] を皮切りに様々に行われているが，中でも園田らによる輸送解析の手法 [2] は流体力学的な視座をもとにしているという点で非常に興味深い．その輸送解析において層数無限大の極限として登場する，確率測度に関する微分方程式が Wasserstein 勾配流である．この微分方程式に関する数値解法は，直感的な反面，計算量の多い変分的解法 [3, 4] が多いが，Benamou&Brenier らの力学系的視点 [5] により差分的な解法も最近いくつか提案されている [6, 7]．さらにその視点を進めた Otto による確率測度空間の形式的な Riemann 構造 [8] に着目し，その構造の離散化を行うと Wasserstein 勾配流に対する構造保存的な空間離散化手法を与えることができる．本稿ではこの手法がある意味で平衡点に高速に収束していることを示し，深層学習との関連を指摘する．

1 緒論

二次モーメントが有限で m 次元 Lebesgue 測度 $\mathrm{d}x$ に関して絶対連続であるような確率測度全体の空間を $P_2(\mathbb{R}^m)$ と表すことにする．本稿では Wasserstein 勾配流と呼ばれる，時間発展する確率測度 $\rho_t\,\mathrm{d}x \in P_2(\mathbb{R}^m)$ に関するある微分方程式

$$\frac{\mathrm{d}\rho_t}{\mathrm{d}t} = -\operatorname{grad}_W \mathcal{F}[\rho] \tag{1}$$

に対して差分解法を考える．ここで $\mathcal{F}$ は適当な汎関数[*1]である．特に $\mathcal{F} = \int_{\mathbb{R}^m} \rho\log\rho$ としたものを Wasserstein 熱方程式と本稿では呼ぶ．grad_W については後ほど説明する．この方程式は最適輸送理論という分野でよく研究されており，ここ十数年の間純粋数学，特に幾何学で盛んに研究されている分野である [10]．一方，最近では機械学習への応用が盛んである．特に，ODENet [1] により端緒が見いだされた深層学習を力学系的に理解する試みにおいて，ここ数年，関連性が指摘されている [11, 12]．本研究の最終ゴールは，この Wasserstein 勾配流に対する数値解析により深層学習の数理に貢献することである．本稿ではその第一歩として，Wasserstein 勾配流に対する空間離散化法を構築することを目的とする．特に，確率測度空間 $P_2(\mathbb{R}^m)$ において成り立つ幾何学的性質に着目し，その構造を離散化することで Wasserstein 勾配流に対する空間離散化手法を導出する[*2]．本稿ではその手法の理論的性質，および数値実験の結果を用いて，園田らにより進められている深層学習の輸送解析 [2] に対し，少し進んだ知見をもたらせたことを報告する．この部分に関しては最後の 4 節で述べることとする．

2 本論

以下内容に入る．大筋としては連続で成り立つ事実を確認し，その離散化を有限グラフを土台として与えるという流れになっている．

2.1 数学的準備と先行研究

以下，可測空間 X に対し，$P(X)$ は X 上の確率測度全体の空間を表す．

[*1] displacement convexity という条件を仮定する．詳細は [9] を参照のこと．

[*2] 導出は異なるが，結果は [13, 14] と一致する．

2.1.1 最適輸送理論と変分的数値解法

以下 $X = \mathbb{R}^m$ とする[*3]．確率測度空間 $P_2(X)$ は線形空間ではない為，$P_2(X)$ 上の微分方程式は関数空間上のそれと同様にはいかない．実際，$P_2(X)$ には以下のように一見直感的でない距離構造が入る．

定義 **2.1.** $\mu, \nu \in P_2(X)$ に対し，$W_2(\mu, \nu)$ を

$$W_2(\mu,\nu) = \sqrt{\inf_{\pi \in \Pi(\mu,\nu)} \int_{X\times X} \|x-y\|^2 \,\mathrm{d}\pi\,(x,y)} \tag{2}$$

で定義する．ここに

$$\Pi(\mu,\nu) = \left\{ \pi \in P(X\times X) \mid \forall A \overset{\mathrm{m'ble}}{\subset} X, \forall B \overset{\mathrm{m'ble}}{\subset} Y, \pi[A\times Y] = \mu[A], \pi[X\times B] = \nu[B] \right\}$$

である．

定理 **2.2** (Kantorovich duality)**.** 式 (2) 中の inf を達成するような最小化元 $\pi \in \Pi(\mu,\nu)$ が存在する．さらに

$$W_2(\mu,\nu)^2 = \max_{\Phi} \left\{ \int_X \varphi \,\mathrm{d}\mu + \int_X \psi \,\mathrm{d}\nu \right\} \tag{3}$$

が成り立つ．ここに

$$\Phi = \left\{ (\varphi,\psi) \in L^1(\mu) \times L^1(\nu) \mid \forall (x,y) \in X\times X, \varphi(x) + \psi(y) \le \|x-y\|^2 \right\}$$

である．

定理 **2.3.** W_2 は $P_2(X)$ 上の距離となる．そこで，W_2 を Wasserstein 距離と呼び，距離空間 $(P_2(X), W_2)$ を Wasserstein 空間と呼ぶ．

定理 **2.4.** $(\mu_k)_{k\in\mathbb{N}}$ を $P_2(X)$ 上の点列とし，$\mu \in P_2(X)$ とする．このとき以下は同値．

1. $W_2(\mu_k, \mu) \to 0$ as $k \to \infty$.
2. $\mu_k \overset{\mathrm{weak}}{\to} \mu$ as $k \to \infty$ であり，かつ，ある $x_0 \in X$ について

$$\lim_{R\to\infty} \limsup_{k\to\infty} \int_{\|x-x_0\|\ge R} \|x-x_0\|^2 \,\mathrm{d}\mu_k = 0$$

が成り立つ．

以上，三定理の証明は [9] に譲る．

式 (2) で確認できる通り，W_2 を求めるためには最小化問題を解く必要がある．このような特徴付けに注目した Wasserstein 勾配流 (1) の解法が次の JKO スキームに代表される変分的数値解法である．

定理 **2.5** (JKO スキーム [3])**.** $\mathcal{F}$ が適当な条件[*4]を満足するとき，$\tau > 0$ に対し次のスキーム[*5]

$$\rho^{k+1} = \operatorname*{arg\,min}_{\rho\in P_2(X)} \left\{ \frac{1}{2\tau} W_2(\rho,\rho^k)^2 + \mathcal{F}[\rho] \right\} \tag{4}$$

で生成される列 $(\rho^k)_{k\in\mathbb{N}}$ は，$\tau \to 0$ の極限で微分方程式 (1) の解に弱収束する．

[*3] 実は大体同じ結果が完備な Riemann 多様体上でも成り立つ．のちにグラフ上に舞台を移すが，グラフは Riemann 多様体の離散化と思え，以下の議論が再現可能である．そのため X と抽象化して書いておく

[*4] ここでは説明しないが displacement convexity 等があれば十分．

[*5] $\mathrm{d}x$ を適当に省略している．

このスキームの収束性は本質的には定理 2.4 によって示される．このように，変分的解法は Wasserstein 距離の有する解析学的に良い性質（定理 2.4）をもとにした，ある意味で直感的な解法となっている．

しかし，JKO スキームを見ればわかるようにこの解法では確率測度に関する最適化問題を各時刻で解く必要があり，計算量の観点からはあまり良い解法とは言えない．Benamou らは定理 2.2 を用いることで確率測度に関する最適化問題を離散凸関数に関する線形不等式制約付き最適化問題に書き換え計算量を削減しようとする試みを行った [4]. しかし，依然計算負荷が大きく実用的な解法とは言えない．

2.1.2 Benamou-Brenier の公式と Otto 解析

前項では Wasserstein 距離の式 (2) による特徴づけを基礎とした変分的数値解法を紹介し，計算量の観点からはあまり良い解法とは言えないことを述べた．このような背景から差分解法を構築しようとする動機付けを得る．勾配流ではない Wasserstein 空間上の微分方程式に関する差分解法は [6, 7, 15, 16] などで与えられている．勾配流に関しては [13] で与えられている[*6]．以上で言及した [6] 以外の先行研究が差分法を構築する基礎としているのが Benamou と Brenier による Wasserstein 距離のもう一つの特徴づけである．

定理 **2.6.** (Banamou-Brenier formula [5])
$\rho_0, \rho_1 \in P_2(X)$ は compact 台を持つとする．このとき

$$W_2(\rho_0,\rho_1)^2 = \inf\left\{\int_0^1 \left(\int_X \rho_t \|v_t\|^2\right) \mathrm{d}t \mid (\rho, v) \in V(\rho_0,\rho_1)\right\} \tag{5}$$

$$= \inf\left\{\int_0^1 \left(\int_X \rho_t \|\boldsymbol{\nabla} u_t\|^2\right) \mathrm{d}t \mid (\rho, \boldsymbol{\nabla} u) \in V(\rho_0,\rho_1)\right\} \tag{6}$$

が成り立つ．ここで $V(\rho_0,\rho_1)$ は以下の条件を満たす $(\rho, v) = (\rho_t, v_t)_{0\le t\le 1}$ の全体の集合である：

$$\rho \in C([0,1]; w* - P_2(X)), \tag{7}$$

$$v \in L^2(\mathrm{d}\rho_t\,\mathrm{d}t), \tag{8}$$

$$\bigcup_{0\le t\le 1} \mathrm{Supp}\,\rho_t; \text{bounded}, \tag{9}$$

$$\frac{\partial \rho_t}{\partial t} + \text{“}\boldsymbol{\nabla}\cdot\text{”}(\rho_t v_t) = 0 \quad \text{in weak sense} \quad \underline{(\text{連続の式})}, \tag{10}$$

$$\rho_{t=0} = \rho_0, \rho_{t=1} = \rho_1. \tag{11}$$

ただし作用素 “$\boldsymbol{\nabla}\cdot$” に関しては後ほど定義 2.7 の中で説明する．定理 2.6 中の $\int_X \rho_t\|v_t\|^2 = \int_X \rho_t\|\boldsymbol{\nabla} u_t\|^2$ を曲線 $(\rho_t)_{0\le t\le 1}$ の “微小長さ” だと思うと，形式的に接空間 $T_\rho P_2(X)$[*7]の元に計量を定めることができる．これが Otto 解析といわれる次の形式的 Riemann 構造（計量）である．

定義 **2.7.** $\left(\frac{\partial\hat\rho_1}{\partial t}\right), \left(\frac{\partial\hat\rho_2}{\partial t}\right) \in T_\rho P_2(X)$ に対し，その間の内積を

$$\left\langle \left(\frac{\partial\hat\rho_1}{\partial t}\right), \left(\frac{\partial\hat\rho_2}{\partial t}\right) \right\rangle_\rho = \int_X \rho \langle \boldsymbol{\nabla}\hat u_1, \boldsymbol{\nabla}\hat u_2\rangle \tag{12}$$

で定める．ここで $u_1, u_2 \in C_c^\infty(X)$[*8]は

$$-\text{“}\boldsymbol{\nabla}\cdot\text{”}(\rho\boldsymbol{\nabla}\hat u_1) = \left(\frac{\partial\hat\rho_1}{\partial t}\right), -\text{“}\boldsymbol{\nabla}\cdot\text{”}(\rho\boldsymbol{\nabla}\hat u_2) = \left(\frac{\partial\hat\rho_2}{\partial t}\right)$$

[*6] 概要を書いた段階ではこの論文を認知しておらず，新しい手法を提案できたと考えていたが，本論文を精査したところ本質的には同じ手法を提案していることが判明した．

[*7] “形式的には"の意味は “基礎的な有限次元多様体論での導入同様 $\rho \in P_2(X)$ を通る $P_2(X)$ 内の曲線 $c : [-\varepsilon,\varepsilon] \to P_2(X)$ を考え，$\dot c(0)$ のことを接空間の元だと思えばよい"ということである．厳密には定理 2.7 を踏まえ，接空間 $T_\rho P_2(X)$ を

$$T_\rho P_2(X) = \overline{\{\boldsymbol{\nabla}\varphi \mid \varphi \in C_c^\infty(X)\}}$$

と定めてしまう [8]．ただし閉包は定理 2.7 で定義される内積から誘導される位相についてとる．ここで $C_c^\infty(X)$ は compact 台を持つ無限回連続微分可能な X 上の実数値関数全体を表す．

[*8] 一意に定まるのであればもっと広い空間でもよいはず．

を満たす．ただし，作用素 “$\boldsymbol{\nabla}\cdot$” については以下の双対性により定義する:

$$\forall\varphi\in C_c^\infty(X),\quad \int_X \varphi\,\mathrm{d}(\text{“}\boldsymbol{\nabla}\cdot\text{”}(\rho v))=-\int_X\langle\boldsymbol{\nabla}\varphi,v\rangle\,\mathrm{d}\rho\,. \tag{13}$$

定義 2.7 により汎関数に対する勾配 grad_W を定義することができるが，実は陽に書き下すこともできる．すなわち次が成り立つ．

定理 **2.8.** $P_2(X)$ 上の汎関数 $\mathcal{F}$ に対し，$\mathrm{grad}_W\mathcal{F}[\rho]\in T_\rho P_2(X)$ を

$$\forall\left(\frac{\partial\hat\rho}{\partial t}\right)\in T_\rho P_2(X),\left\langle\mathrm{grad}_W\mathcal{F}[\rho],\left(\frac{\partial\hat\rho}{\partial t}\right)\right\rangle_\rho=D\mathcal{F}[\rho]\left(\left(\frac{\partial\hat\rho}{\partial t}\right)\right)$$

で定義する．十分 $\mathcal{F}$ が滑らかであるとき，

$$\mathrm{grad}_W\mathcal{F}[\rho]=-\text{“}\boldsymbol{\nabla}\cdot\text{”}\left(\rho\boldsymbol{\nabla}\frac{\delta\mathcal{F}}{\delta\rho}\right) \tag{14}$$

が成り立つ．ここで $\frac{\delta\mathcal{F}}{\delta\rho}$ は L^2 内積に対する変分であり，特に $\mathcal{F}[\rho]=\int_X F(\rho)$ であるとき $\frac{\delta\mathcal{F}}{\delta\rho}=F'(\rho)$ である．

証明. $\mathcal{F}[\rho]=\int_X F(\rho)$ の場合のみ示す．$\mathrm{grad}_W\mathcal{F}[\rho]$ に対する連続の式の解を f とおく．すなわち

$$\mathrm{grad}_W\mathcal{F}[\rho]+\text{“}\boldsymbol{\nabla}\cdot\text{”}(\rho\boldsymbol{\nabla}f)=0$$

とする．$f=\frac{\delta\mathcal{F}}{\delta\rho}$ を示せばよい．

任意に $\left(\frac{\partial\hat\rho}{\partial t}\right)\in T_\rho P_2(X)$ をとり，それと連続の式により対応する関数を φ とおく．定義 2.7 から

$$\begin{aligned}\left\langle\mathrm{grad}_W\mathcal{F}[\rho],\left(\frac{\partial\hat\rho}{\partial t}\right)\right\rangle_\rho&=\frac{\mathrm{d}}{\mathrm{d}\epsilon}\mathcal{F}\left[\rho+\epsilon\left(\frac{\partial\hat\rho}{\partial t}\right)\right]\Big|_{\epsilon=0}\\&=\int_X\frac{\mathrm{d}}{\mathrm{d}\epsilon}F\left(\rho+\epsilon\left(\frac{\partial\hat\rho}{\partial t}\right)\right)\Big|_{\epsilon=0}\\&=\int_x\frac{\delta\mathcal{F}}{\delta\rho}\left(\frac{\partial\hat\rho}{\partial t}\right)\end{aligned}$$

が成り立つ．一方，

$$\begin{aligned}\left\langle\mathrm{grad}_W\mathcal{F}[\rho],\left(\frac{\partial\hat\rho}{\partial t}\right)\right\rangle_\rho&=\int_X\rho\langle\boldsymbol{\nabla}f,\boldsymbol{\nabla}\varphi\rangle\\&=\int_X\langle\boldsymbol{\nabla}f,\rho\boldsymbol{\nabla}\varphi\rangle\\&=\int_X f(-\text{“}\boldsymbol{\nabla}\cdot\text{”}(\rho\boldsymbol{\nabla}\varphi))\\&=\int_X f\left(\frac{\partial\hat\rho}{\partial t}\right)\end{aligned}$$

である．$\left(\frac{\partial\hat\rho}{\partial t}\right)\in T_\rho P_2(X)$ は任意にとったので，主張を得る． □

次節においては上記の特徴づけをもとに Wasserstein 空間を離散化していく．

2.2 グラフ上の Otto 解析

前項までで説明した Wasserstein 空間に備わっている幾何的構造を離散化することにより，Wasserstein 勾配流に対する差分法を構築していく．なお，以降の議論は導入が異なるものの，出てくる結果は [13] と同じものであった．

以下では Wasserstein 空間の離散化として，X が重み付き有限無向グラフ $G=(V,E,\Omega)$ である場合を考える．ここで頂点集合を $V=\{i\}_{i=1}^{\#V}$，辺集合を $E=\{e_k=[ij]\}_{k=1}^{\#E}$，向きを区別した辺集合[*9]を $\tilde E\subset V^2$，辺上の重みを $\Omega=\{\omega_{ij}\}$ と書き，G は連結で自己ループがないものとする．また N_i を i の隣接ノードの集合とする．

[*9] グラフの辺集合 E 自体は無向だが，関数空間 $\mathcal{E}$ を定義するためあえて適当に向き付けした頂点集合も用意する．

まずグラフ上に L^2 空間（関数空間）を構築する．以下は [17] を参考にした．

定義 **2.9.** グラフ $G=(V,E,\Omega)$ に対し，

$$\mathcal{V}=\{u:V\to\mathbb{R}\},\tag{15}$$

$$\mathcal{E}=\left\{\phi:\tilde{E}\to\mathbb{R}\mid\forall[ij]\in E,\phi_{ij}=-\phi_{ji}\right\}\tag{16}$$

とし，$u,v\in\mathcal{V},\phi,\psi\in\mathcal{E}$ に対し種々の演算を以下で定義する：

$$\langle u,v\rangle_{\mathcal{V}}=\sum_{i\in V}u_iv_i,\tag{17}$$

$$\langle\phi,\psi\rangle_{\mathcal{E}}=\sum_{[ij]\in E}\phi_{ij}\psi_{ij}=\frac{1}{2}\sum_{ij\in\tilde{E}}\phi_{ij}\psi_{ij},\tag{18}$$

$$(\phi\cdot\psi)_i=\frac{1}{2}\sum_{j\in N_i}\phi_{ij}\psi_{ij},\tag{19}$$

$$(\boldsymbol{\nabla}u)_{ij}=\sqrt{\omega_{ij}}(u_j-u_i).\tag{20}$$

次に Wasserstein 空間 $P(G)$ を構築する．$P(G)$ の元はデルタ測度の重み付き和として表す．つまり

$$P(G)=\left\{\rho=\sum_{i\in V}\rho_i\delta_i\mid\sum_{i\in V}\rho_i=1\ \&\ \forall i\in V,\rho_i>0\right\}\tag{21}$$

に Wasserstein 距離 W_2 で距離構造を入れたものをグラフ上の Wasserstein 空間とする．$\rho_i\neq0$ としたのは後の都合である．

さらに点 $\rho\in P(G)$ に対し接空間[*10] $T_\rho P(G)$ を次で定義する．

定義 **2.10.** $\rho=\sum_{i\in V}\rho_i\delta_i\in P(G)$ の接空間 $T_\rho P(G)$ を

$$T_\rho P(G)=\{-\text{“}\boldsymbol{\nabla}\cdot\text{”}(\rho\boldsymbol{\nabla}\varphi)\mid\varphi\in\mathcal{V}\}\tag{22}$$

で定義する．ここで，作用素 “$\boldsymbol{\nabla}\cdot$”$(\rho\cdot):\mathcal{E}\ni v\mapsto T_\rho P(G)\in$ “$\boldsymbol{\nabla}\cdot$”(ρv) は以下が成り立つように定義する．

$$\forall f\in\mathcal{V},\sum_{i\in V}f_i(\text{“}\boldsymbol{\nabla}\cdot\text{”}(\rho v))_i=-\sum_{i\in V}(\boldsymbol{\nabla}f\cdot v)_i\rho_i\tag{23}$$

実は “$\boldsymbol{\nabla}\cdot$” は次のように閉じた形で表すことができる[*11]．

定理 **2.11.** $\rho\in P(G),v\in\mathcal{E}$ について

$$\forall i\in V,\quad(\text{“}\boldsymbol{\nabla}\cdot\text{”}(\rho v))_i=\sum_{j\in N_i}\sqrt{\omega_{ij}}\frac{\rho_i+\rho_j}{2}v_{ij}\tag{24}$$

が成り立つ．

[*10] $P(G)$ は既にユークリッド空間内の部分多様体と見做せ，接空間が自動的に定義されるが，今回は計量としてユークリッド計量ではないものを定義したいのでわざわざ “接空間” を定義している．

[*11] この部分は本質的には [13] で既に発見されている．

証明.

$$\begin{aligned}
\sum_{i\in V} f_i(\text{“}\boldsymbol{\nabla}\cdot\text{”}(\rho v))_i &= -\sum_{i\in V}(\boldsymbol{\nabla} f\cdot v)_i\rho_i \\
&= -\sum_{i\in V}\frac{1}{2}\sum_{j\in N_i}\sqrt{\omega_{ij}}(f_j-f_i)v_{ij}\rho_i \\
&= \sum_{[ij]\in E}\sqrt{\omega_{ij}}(f_i-f_j)v_{ij}\rho_i \\
&= \sum_{[ij]\in E}\sqrt{\omega_{ij}}f_iv_{ij}\rho_i - \sum_{[ij]\in E}\sqrt{\omega_{ij}}f_jv_{ij}\rho_i \\
&= \sum_{[ij]\in E}\sqrt{\omega_{ij}}f_iv_{ij}\rho_i + \sum_{[ij]\in E}\sqrt{\omega_{ij}}f_iv_{ij}\rho_j \\
&= \sum_{[ij]\in E}\sqrt{\omega_{ij}}f_iv_{ij}(\rho_i+\rho_j) \\
&= \frac{1}{2}\sum_{i\in V}f_i\sum_{j\in N_i}\sqrt{\omega_{ij}}v_{ij}(\rho_i+\rho_j) \\
&= \sum_{i\in V}f_i\left(\sum_{j\in N_i}\sqrt{\omega_{ij}}\frac{\rho_i+\rho_j}{2}v_{ij}\right)
\end{aligned}$$

から主張を得る. □

今示した定理により接空間 $T_\rho P(G)$ の元が陽に表示できるようになった.

よって，$\mathbb{R}^m$ 上の Otto 解析 (2.7) から，グラフ上の Otto 解析は次のように定めることができる.

定義 **2.12.** $\left(\frac{\partial\hat{\rho}_1}{\partial t}\right), \left(\frac{\partial\hat{\rho}_2}{\partial t}\right) \in T_\rho P(G)$ に対し，その間の内積を

$$\left\langle \left(\frac{\partial\hat{\rho}_1}{\partial t}\right), \left(\frac{\partial\hat{\rho}_2}{\partial t}\right)\right\rangle_\rho = \sum_{i\in V}\rho_i\langle\boldsymbol{\nabla}\hat{u}_1, \boldsymbol{\nabla}\hat{u}_2\rangle_{\mathcal{E}} \tag{25}$$

で定める．ここで $u_1, u_2 \in \mathcal{V}$ は

$$\forall i\in V, -\sum_{j\in N_i}\omega_{ij}\frac{\rho_i+\rho_j}{2}((u_1)_j-(u_1)_i) = \left(\frac{\partial\hat{\rho}_1}{\partial t}\right), -\sum_{j\in N_i}\omega_{ij}\frac{\rho_i+\rho_j}{2}((u_2)_j-(u_2)_i) = \left(\frac{\partial\hat{\rho}_2}{\partial t}\right)$$

を満たす.

勾配 grad_W も定理 2.8 と同様の証明により陽に表すことができる.

定理 **2.13.** $P(G)$ 上の汎関数 $\mathcal{F} = \sum_{i\in V}F(\rho_i)$ に対し，$\mathrm{grad}_W\mathcal{F}[\rho] \in T_\rho P(G)$ は

$$(\mathrm{grad}_W\mathcal{F}[\rho])_i = -\sum_{j\in N_i}\omega_{ij}\frac{\rho_i+\rho_j}{2}(F'(\rho_j)-F'(\rho_i)) \tag{26}$$

と表せる.

2.3 Wasserstein 熱方程式と線形熱方程式

以下重みは均等に $\omega_{ij}=1$ であるとする．前項までで導出した結果を用いて $\mathcal{F}[\rho]=\mathcal{H}[\rho] := \sum_{i\in V}\rho_i\log\rho_i$ であるときの Wasserstein 勾配流 (1) を考える．これは Wasserstein 熱方程式と呼ばれるのであった．$\mathcal{H}$ はボルツマンエントロピーと呼ばれる．定理 2.13 を用いると，常微分方程式

$$\frac{\mathrm{d}\rho_i(t)}{\mathrm{d}t} = \sum_{j\in N_i}\frac{\rho_i+\rho_j}{2}\log\frac{\rho_j}{\rho_i}, \quad i\in V \tag{27}$$

が導出される．しかし，Chow らが 2012 年に提案した離散化 [18][*12]，同じことだが，関数に関する熱方程式[*13]の空間離散化は，ここだけの記法として，$\rho = (\rho_1, \cdots, \rho_{\#V})^\top$ とすると

$$\frac{\mathrm{d}\rho}{\mathrm{d}t} = -L\rho \tag{28}$$

という常微分方程式になり，式 (27) とは異なる．ここで，L は G のグラフラプラシアンである．そこで，式 (28) を線形の熱方程式と呼び，それとは区別して，式 (27) をグラフ G 上の Wasserstein 熱方程式と呼ぶことにする．以後しばらく “グラフ G 上の"を省略する．

実は式 (28) は式 (27) の平衡点[*14]$(\#V^{-1}, \ldots, \#V^{-1})^\top$ 周りの線形化になっていることが計算により分かる．言い換えると Wasserstein 熱方程式は線形の熱方程式の非線形化であるともいえる．

この非線形作用により，Wasserstein 熱方程式は線形の熱方程式より速い拡散効果を持つことが理論的に分かる．すなわち，線形の熱方程式 (28) もボルツマンエントロピーを減少させるが，減少具合は Wasserstein 熱方程式 (27) の方が真に大きい．この事実のため，本稿では式 (27) を “Wasserstein” 熱方程式と呼んでいる．

定理 **2.14.** $\rho_0 \in P(G)$ を初期値として，$\rho_{\text{Wass.}}(t) = (\rho_{\text{Wass.}i})_{i\in V}$ を常微分方程式 (27) の解，$\rho_{\text{Lin.}}(t) = (\rho_{\text{Lin.}i})_{i\in V}$ を常微分方程式 (28) の解とする．このとき，任意の $t \geq 0$ に対し

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{H}[\rho_{\text{Wass.}}(t)] \leq \frac{\mathrm{d}}{\mathrm{d}t}\mathcal{H}[\rho_{\text{Lin.}}(t)] \leq 0 \tag{29}$$

が成り立つ．ただし等号成立は $\rho_{\text{Wass.}}(t) = \rho_{\text{Lin.}}(t) = (\#V^{-1}, \ldots, \#V^{-1})^\top$ のときに限る．

証明. $\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{H}$ を計算すると，

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{H}[\rho_{\text{Wass.}}(t)] = -\sum_{[ij]\in E} \frac{\rho_i + \rho_j}{2}\left(\log\frac{\rho_j}{\rho_i}\right)^2 \leq 0,$$
$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{H}[\rho_{\text{Lin.}}(t)] = -\sum_{[ij]\in E} (\rho_j - \rho_i)\left(\log\frac{\rho_j}{\rho_i}\right) \leq 0$$

のようになる．さらに，任意の $t > 0$ に対し

$$\log t\left(\frac{t+1}{2}\log t - t + 1\right) \geq 0$$

が成り立つことから

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{H}[\rho_{\text{Wass.}}(t)] - \frac{\mathrm{d}}{\mathrm{d}t}\mathcal{H}[\rho_{\text{Lin.}}(t)] = -\sum_{[ij]\in E} \rho_i \log\frac{\rho_j}{\rho_i}\left(\frac{\frac{\rho_j}{\rho_i}+1}{2}\log\frac{\rho_j}{\rho_i} - \frac{\rho_j}{\rho_i} + 1\right) \leq 0$$

であり，主張を得る． □

3 数値実験

定理 2.14 で Wasserstein 熱方程式がボルツマンエントロピーを線形の熱方程式より真に速く散逸させることが保証された．この結果は時間については連続な場合しか保証していないが，時間を適当に離散化してもこ のような性質がある程度保たれることが予想される．そこで以下では数値実験によりそれを確認してみる．

[*12] しかし 2018 年に Chow らは式 (27) と同じ手法を導出し [13]，事実上 2012 年の結果 [18] を訂正している．
[*13] 例えば $\mathbb{R}$ 値の関数に関する熱方程式 $\partial_t u = \Delta u = \partial_{xx} u$ を指す．これを以下 “通常の"熱方程式と呼ぶことがある．
[*14] 時間発展する微分方程式 $\frac{\mathrm{d}y}{\mathrm{d}t} = f(y)$ のついて $f(y) = 0$ となるような定数解 y を一般に平衡点という．

3.1　設定

図 1 で与えられるグラフ $G = (V, E)$ 上で，座標 $(0.5, 0.5)$ にある点 $c \in V$ については $\rho_c = 1 - (500 - 1) \cdot 10^{-10}$，$i \in V \setminus \{c\}$ に対しては $\rho_i = 10^{-10}$ であるような初期値を与えた Wasserstein 熱方程式 (27) と線形の熱方程式 (28) を，Python の常微分方程式求積ライブラリの一つである `scipy.integrate.odeint`[*15]を用いて時間刻み幅 $\tau = 0.01$ で数値的に解く．

ただしグラフ G は $\#V = 500$ かつ半径 0.1 以内の頂点としか辺結合していないようなものの中からランダムで作成されたものである．

3.2　結果

頂点 $i \in V$ の半径を ρ_i に比例させて結果を描画した（図 1）．左列が Wasserstein 熱方程式，右列が線形の熱方程式を解いて結果であり下に行くほど時間発展していくが，左の列の方が中心から離れた点にも赤色の頂点が確認できる時刻が早い，つまり Wasserstein 熱方程式の方が拡散効果が大きいことが確認できる．

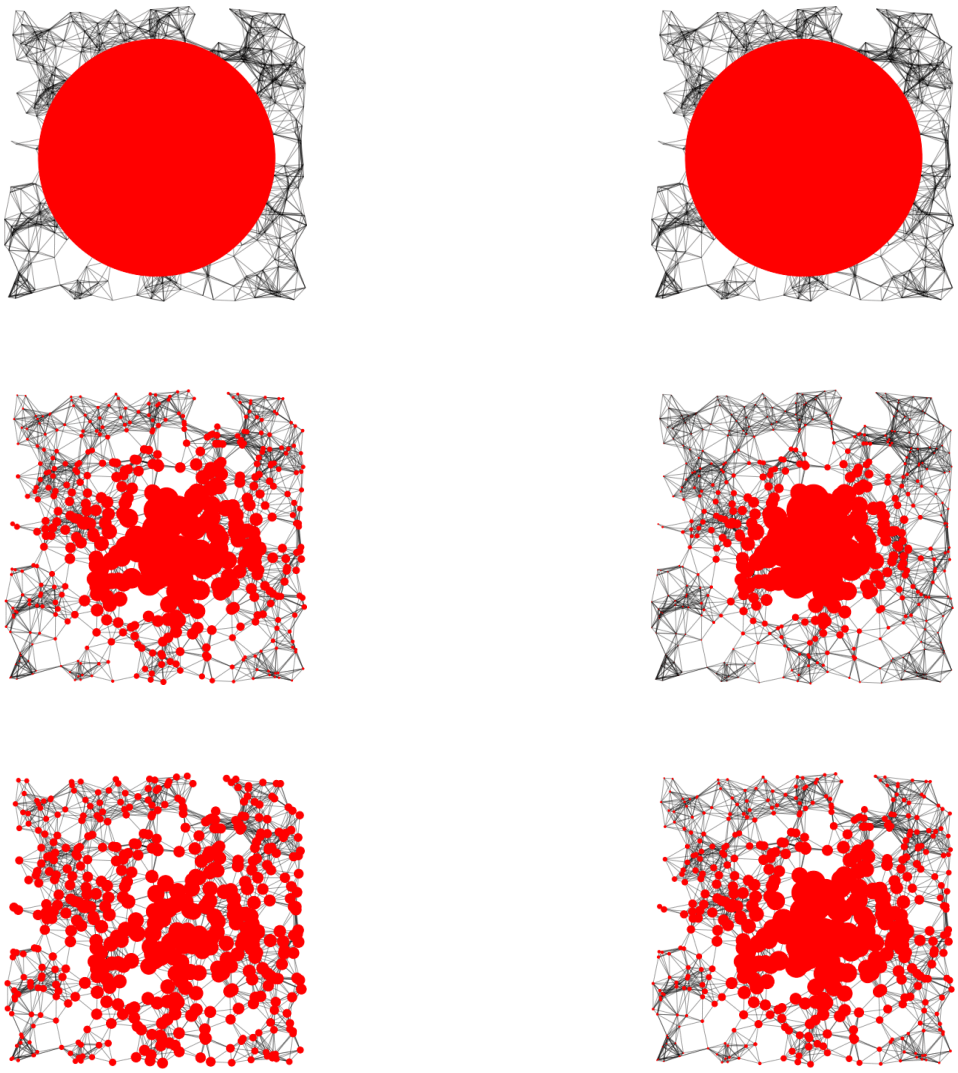

図 1　Adams 法で $\#V = 500$, 時間刻み $\tau = 0.01$，初期データとして中心点以外は $\rho_i = 10^{-10}$ を与え，解いた結果．左列が Wasserstein 熱方程式 (27)，右列が線形の熱方程式 (28) を表し，行は時刻を表し，一番上から $t = 0.00, 0.02, 0.06, 0.10$ に対応する．値 ρ_i の大きさが赤点の半径に比例するように描画した．一番下の行の左右を比べると左の Wasserstein 熱方程式の方が中心から遠い点の半径が大きい．

[*15] つまり Adams-Bashforth-Moulton 法

ボルツマンエントロピー $\mathcal{H}$ の時間発展も確認してみたところ，図 2 のようになった．

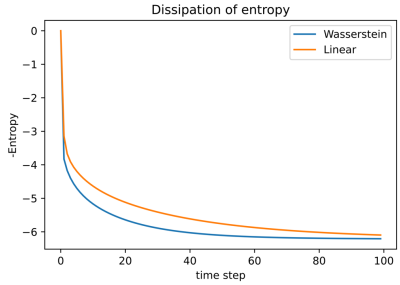

図 2　図 1 と同じ条件で，ボルツマンエントロピー $\mathcal{H}$ が時間が経つに伴って減少することを確認した．

実は、熱方程式以外の場合でも Wasserstein 勾配流の平衡点への収束の速さが確認できる．

4　深層学習への応用

定理 2.14 や数値実験で示された Wassersteinn 熱方程式の平衡点への収束の速さ，エントロピーの減少の速さは一体何を示唆しているのであろうか．ここで，園田らの深層学習に対する輸送解析 [2] の結果から考察を試みる．以下，Denoising Autoencoder(DAE) と呼ばれる，訓練データにわざとノイズを加え，それ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を考える．園田らによると，DAE を多層にし，さらにその層数無限大の極限を考えると，学習器 $\boldsymbol{g}_t$[*16]と，学習器 $\boldsymbol{g}_t$ により予測されたデータの分布 π_t は以下を満たすことが指摘されている．

$$\frac{\partial \boldsymbol{g}_t}{\partial t} = \boldsymbol{\nabla} \log \pi_t(\boldsymbol{x}) \tag{30}$$

$$\frac{\partial \pi_t}{\partial t} = -\text{“}\Delta\text{”}\pi_t(\boldsymbol{x}) \tag{31}$$

式 (31) は Wasserstein 熱方程式の時間パラメータ t を $-t$ にしたものであるので，時間逆向き Wasserstein 熱方程式と呼ぶ．この逆向きの方程式は，順向きの逆問題，すなわち，Wasserstein 熱方程式の時刻 T の値 $\rho(T)$ から初期値 $\rho(0)$ の値を推定する問題を解いていることに対応する．しかし，通常の熱方程式に対する逆問題は安定性がない，つまり，観測データ $\rho(T)$ に少しでも誤差が入ると推定値 $\rho(0)$ に著しい影響が出てしまう．それは通常の熱方程式がもつ平滑化作用，大雑把にいえば平衡点への収束の速さが引き起こす現象である．この事実に鑑みると，（通常の熱方程式の離散化である）線形の熱方程式より収束が速い[*17]，Wasserstein 熱方程式の逆問題たる深層 DAE は，“安定性がより一層ない” 逆問題を解いていることになる．しかし，実際の DAE がよい学習結果を出力することを考えると，深層学習において使われる種々の正則化手法が学習の安定化に非常に効果的に作用しているということが予想される．

5　結論

上の考察により，少なくとも Deep DAE については，学習器 $\boldsymbol{g}_t$ の振る舞いを解析することは，学習器によって輸送される分布 π_t が従う Wasserstein 勾配流 (31) の性質を調べることに帰着できる．一般の深層学習モデルにおいてこのような議論ができるかは未だ on going な課題である．しかし，本稿では述べないが，パラメータ数無限大（幅無限大）の極限において Wasserstein 勾配流が登場する例も知られており [11, 12]，深層学習理論の解明のために Wasserstein 勾配流の数学的性質を調べることは今後益々必要になる．

さらに，Wasserstein 勾配流を数値解析学的に調べることも重要である．なぜならば，上の例で見た通り実際の学習器は時間に関して離散化されており，取りうるデータも離散的な場合がある．連続な偏微分方程式を離散化した場合に様々な構造が破壊され得ることは古典的に数値解析学でよく知られた事実であり，層無限大の理想的な学習器を

[*16] ここで時刻 t は層数だと思えばよい．
[*17] 時間逆向きでは定理 2.14 の不等号の向きがすべて逆向きになることに注意

離散化した場合にどのような性質が保たれ，どのような性質が壊れるのか，性質を壊さないためにはどのように離散化すればよいのか，というのはこれから中長期的に取り組まねばならない課題であると考えられる．

参考文献

[1] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud. “Neural Ordinary Differential Equations”. In: *Advances in Neural Information Processing Systems 31*. 2018, pp. 6571–6583.

[2] S. Sonoda and N. Murata. “Transport Analysis of Infinitely Deep Neural Network”. In: *Journal of Machine Learning Research* 20.2 (2019), pp. 1–52.

[3] R. Jordan, D. Kinderlehrer, and F. Otto. “The Variational Formulation of the Fokker–Planck Equation”. In: *SIAM Journal on Mathematical Analysis* 29.1 (1998), pp. 1–17.

[4] J.-D. Benamou, G. Carlier, Q. Mérigot, and É. Oudet. “Discretization of functionals involving the Monge–Ampère operator”. In: *Numerische Mathematik* 134.3 (2016), pp. 611–636.

[5] J.-D. Benamou and Y. Brenier. “A numerical method for the optimal time-continuous mass transport problem and related problems”. In: *Monge Ampère Equation: Applications to Geometry and Optimization* (Deerfield Beach, FL, 1997). Vol. 226. Contemporary Mathematics. Providence, R.I: American Mathematical Society, 1999, pp. 1–11.

[6] E. Carlini and F. J. Silva. *Numerical methods for non-linear Fokker Planck equations and applications to Mean Field Games*. Workshop III: Mean Field Games and Applications Part of the Long Program High Dimensional Hamilton-Jacobi PDEs. 2020.

[7] J. Cui, L. Dieci, and H. Zhou. *Time Discretizations of Wasserstein-Hamiltonian Flows*. 2020. arXiv: 2006.09187 [math.NA].

[8] F. Otto. “The geometry of dissipative evolution equations: the porous medium equation”. In: *Communications in Partial Differential Equations* 26.1-2 (2001), pp. 101–174.

[9] C. Villani. *Topics in Optimal Transportation (Graduate Studies in Mathematics, Vol. 58)*. American Mathematical Society, 2003.

[10] S. Ohta and A. Takatsu. “Equality in the logarithmic Sobolev inequality”. In: *manuscripta mathematica* 162.1 (2020), pp. 271–282.

[11] A. Nitanda and T. Suzuki. *Stochastic Particle Gradient Descent for Infinite Ensembles*. 2017. arXiv: 1712.05438 [stat.ML].

[12] L. Chizat and F. Bach. “On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport”. In: *Advances in Neural Information Processing Systems 31*. 2018, pp. 3036–3046.

[13] H. Z. Shui-Nee Chow Wuchen Li. “Entropy dissipation of Fokker-Planck equations on graphs”. In: *Discrete & Continuous Dynamical Systems - A* 38 (2018), p. 4929.

[14] S.-N. Chow, L. Dieci, W. Li, and H. Zhou. “Entropy dissipation semi-discretization schemes for Fokker–Planck equations”. In: *Journal of Dynamics and Differential Equations* 31.2 (2019), pp. 765–792.

[15] N. Loy and M. Zanella. *Structure preserving schemes for nonlinear Fokker-Planck equations with anisotropic diffusion*. 2020. arXiv: 1905.02970 [math.NA].

[16] S.-N. Chow, W. Li, and H. Zhou. “A discrete Schrödinger equation via optimal transport on graphs”. In: *Journal of Functional Analysis* 276.8 (2019), pp. 2440–2469.

[17] Y. van Gennip, N. Guillen, B. Osting, and A. L. Bertozzi. “Mean Curvature, Threshold Dynamics, and Phase Field Theory on Finite Graphs”. In: *Milan Journal of Mathematics* 82.1 (2014), pp. 3–65.

[18] S.-N. Chow, W. Huang, Y. Li, and H. Zhou. “Fokker–Planck Equations for a Free Energy Functional or Markov Process on a Graph”. In: *Archive for Rational Mechanics and Analysis* 203.3 (2012), pp. 969–1008.