Wasserstein 距離とはなにか？

2024 年 6 月 6 日

吉村 耕平

最適輸送理論とは、分布間の距離を輸送のコストによって与え、その結果顕現する幾何学構造を調べる学問であり、特に Wasserstein 距離と呼ばれる一連の距離関数が主役となる。昨今物理学においても、非平衡系の熱力学との関係から注目を集めている。最適輸送理論に関する入門的テキストは数多く存在するが、いずれも数学者の言葉遣いで書かれたものであり、管見の限りでは物理学者が親しみやすい言葉で書かれたものは存在しなかった[1]。本ノートでは、数学的な厳密性をあえて犠牲にしつつ、普段筆者がどういう気持ちで最適輸送理論を理解しているかを、時にざっくばらんな表現も交えつつ伝えることを目指した。測度論に特有の記法や言葉遣いは可能な限り排し、物理学科の学生なら理解できるレベルになるような気をつけた。本ノートは本来、沙川 ERATO 岡田新学術合同勉強会のために書かれたものであるが、人々が最適輸送理論への理解を深める一助になればと願い、拙いながらも公開するところである。

目次

最新版は https://ykohei.com/files.html.

[1] 佐藤竜馬氏の書かれた本 [1] は、応用数学的な立場からでありながら極めてわかりやすく書かれている。

はじめに：何を考えるか？

一般に、測定や制御は不完全であるから、古典量子問わず物理を考える上で確率の概念を用いることが有用になる。その際、状態はもはや決定論的な変数でなく、ある確率分布 p に従う確率変数によって特徴づけられることだろう。あるいは、確率分布によって物理状態は表されると言っても良い。

ときに、熱力学はある操作の実現可能性をエントロピーが増加するか否かによって与える。これは非平衡系でも通用すると信じられており、ある操作を行うためにはエントロピーが増大することが要請される。さらに、特定の枠組みのもとではより定量的な予言ができて、実現したい「機能の量」（たとえば精度、スピード）を実現するためには、一定量以上のエントロピーの増大が必要である、と言ったことが示される。このような関係を最も一般的に論じようとしたとき役に立つのが、状態の空間における距離の概念である。

確率分布によって表される状態の変化を測るためには、確率分布間の距離が定義できれば十分である。そのためのもっとも素朴な方法が、L^1-距離

$$L(p,q) = \sum_x |p_x - q_x| \tag{1}$$

である。もっと解析的な性質がよく、しかも奇跡的に Brown 運動のようなダイナミクスの熱力学理論と相性が良かったのが、2-Wasserstein 距離と呼ばれる距離であった。

本ノートでは、2-Wasserstein を含む、より一般の設定において連続分布間の Wasserstein 距離の基本的な性質を説明する（1 章）とともに、離散自由度系特有の性質に着目した場合の Wasserstein 距離の振る舞いについても議論する（2,3 章）。なお、筆者の守備範囲の問題により、本ノートは数理的な話題に終始しており、実用上重要な計算アルゴリズムに関する話は一切行われていない。

1　連続系における Wasserstein 距離

この章では連続変数を考える。ある変数が値 $\boldsymbol{x} \in X = \mathbb{R}^N$ を取る確率を、確率分布関数 $p(\boldsymbol{x})$ によって与える。確率分布関数は非負性 $p(\boldsymbol{x}) \geq 0$ および規格化 $\int_X p(\boldsymbol{x})dx = 1$ を満たす。

1.1　輸送コストとして

まず初めに最も標準的な Wasserstein 距離の定め方を説明し、その性質を示す。以下では常に初期分布 $p^{(0)}$ と終分布 $p^{(1)}$ の間の距離を考える。

図 1 のように、確率分布 p を砂山（＝砂の量の分布）とみなし、$p^{(0)}$ を別の形の砂山 $p^{(1)}$ に変化させることを考えよう。このときの「コスト」が $p^{(0)}$ と $p^{(1)}$ の間の距離を与える。「位置」$\boldsymbol{x} \in X$ から $\boldsymbol{y} \in X$ に砂を $\pi(\boldsymbol{x},\boldsymbol{y})$ だけ動かすとする。行き先を全て足し上げれば、$\boldsymbol{x}$ にもともとあった砂の量 $p^{(0)}(\boldsymbol{x})$ に一致しなければならないから、次が要請される：

$$\int \pi(\boldsymbol{x},\boldsymbol{y})dy = p^{(0)}(\boldsymbol{x}). \tag{2}$$

逆に、行き先 $\boldsymbol{y}$ に運ばれてくる砂の量は、目的の量 $p^{(1)}(\boldsymbol{y})$ と合致する必要がある。つまり

$$\int \pi(\boldsymbol{x},\boldsymbol{y})dx = p^{(1)}(\boldsymbol{y}) \tag{3}$$

が成り立つはずである。これら 2 条件に加えて、非負性 $\pi(\boldsymbol{x},\boldsymbol{y}) \geq 0$ を満たす π を**カップリング**と呼び、$p^{(0)}$, $p^{(1)}$ 間のカップリングの集合を $\Pi(p^{(0)}, p^{(1)})$ と書くことにする。たとえば、$\pi(\boldsymbol{x},\boldsymbol{y}) = p^{(0)}(\boldsymbol{x})p^{(1)}(\boldsymbol{y})$ は最も簡単なカップリングの例である。このカップリングは、すべての場所 $\boldsymbol{x}$ から $\boldsymbol{y}$ へ、重み $p^{(1)}(\boldsymbol{y})$ に比例する分だけ輸送することを意味する。

次に、砂を運ぶためのコストを導入する。$\boldsymbol{x}$ から $\boldsymbol{y}$ へ単位量の砂を運ぶためのコストを $c(\boldsymbol{x},\boldsymbol{y})$ とする。このとき、

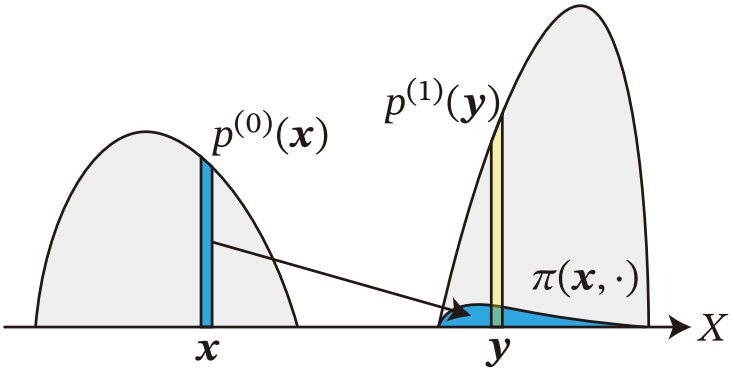

図 1　カップリングの例。各 $\boldsymbol{x}$ の重み $p^{(0)}(\boldsymbol{x})$ は $\pi(\boldsymbol{x},\cdot)$ として分配される。$\pi(\boldsymbol{x},\boldsymbol{y})$ は、$p^{(1)}(\boldsymbol{y})$ のうちの一部を与える（青い部分と黄色い部分の重なり）。$\boldsymbol{x}$ について足し合わせれば $p^{(1)}(\boldsymbol{y})$ を与えるし、$\boldsymbol{y}$ について足し合わせれば $p^{(0)}(\boldsymbol{x})$ を与える。

カップリング π を用いた場合の総コストは次で与えられる：

$$C[\pi] = \int c(\boldsymbol{x},\boldsymbol{y})\pi(\boldsymbol{x},\boldsymbol{y})d\boldsymbol{x}d\boldsymbol{y}. \tag{4}$$

砂山の比喩から言えば、コスト関数 $c(\boldsymbol{x},\boldsymbol{y})$ はガソリン代ないし人件費のようなものである。そこで、X 上の距離関数 $d(\boldsymbol{x},\boldsymbol{y})$ の増加関数をコスト関数に選ぶことが考えられる。距離関数とは、任意の $\boldsymbol{x},\boldsymbol{y},\boldsymbol{z} \in X$ に対して、次の 4 条件を満たすもののことである：

$$\begin{aligned}
&\text{1. 非負性} && d(\boldsymbol{x},\boldsymbol{y}) \geq 0,\\
&\text{2. 非退化性} && d(\boldsymbol{x},\boldsymbol{y}) = 0 \iff \boldsymbol{x} = \boldsymbol{y},\\
&\text{3. 対称性} && d(\boldsymbol{x},\boldsymbol{y}) = d(\boldsymbol{y},\boldsymbol{x}),\\
&\text{4. 三角不等式} && d(\boldsymbol{x},\boldsymbol{y}) + d(\boldsymbol{y},\boldsymbol{z}) \geq d(\boldsymbol{x},\boldsymbol{z}).
\end{aligned} \tag{5}$$

Euclid 距離 $d(\boldsymbol{x},\boldsymbol{y}) = [\sum_{i=1}^{N} |x_i - y_i|^2]^{1/2}$ が最も典型的な選択であるが、X が曲がった空間だったらその空間上の Riemann 距離を選ぶことが自然になるだろう。X が離散的な場合、後で見るように離散集合上の結合関係に着目した距離の定義も可能になる。

$c(\boldsymbol{x},\boldsymbol{y}) = d(\boldsymbol{x},\boldsymbol{y})^r$ $(1 \leq r < \infty)$ の場合の総コストを $C_r[\pi]$ とすれば、$(C_r[\pi])^{1/r}$ をカップリング π について最小化したものが r-Wasserstein 距離と呼ばれるものになる。すなわち、確率分布 $p^{(0)}$ および $p^{(1)}$ の間の **r-Wasserstein 距離** $\mathcal{W}_r(p^{(0)}, p^{(1)})$ は次で定義される：

$$\mathcal{W}_r(p^{(0)}, p^{(1)}) \coloneqq \inf_{\pi \in \Pi(p^{(0)}, p^{(1)})} \left[\int d(\boldsymbol{x},\boldsymbol{y})^r \pi(\boldsymbol{x},\boldsymbol{y}) d\boldsymbol{x}d\boldsymbol{y} \right]^{1/r}. \tag{6}$$

この最小化問題には、かなり一般的に最適解が存在することが知られている[2]。連続系で最もよく用いられるのが $r = 2$ の場合であるが、そのようなパラボリックな重み付け自体に物理的必然性があるわけでなはい。実際、これを離散系に拡張しようとした場合、$r = 2$ では不都合なことが発生する（後述）。

Wasserstein 距離は最小化によって定義されるため、任意のカップリング $\pi' \in \Pi(p^{(0)}, p^{(1)})$ に関して不等式

$$\mathcal{W}_r(p^{(0)}, p^{(1)}) \leq \left[\int d(\boldsymbol{x},\boldsymbol{y})^r \pi'(\boldsymbol{x},\boldsymbol{y}) d\boldsymbol{x}d\boldsymbol{y} \right]^{1/r} \tag{7}$$

が成り立つ。

[2] [2, Thm.4.1]. 上半連続（不連続な点の値が、大きい側に一致する）関数 a, b で $p^{(0)}, p^{(1)}$ でそれぞれ絶対値を積分した時有限なものによって、下半連続（不連続な点の値が小さい側に一致する）関数 c が $c(\boldsymbol{x},\boldsymbol{y}) \geq a(\boldsymbol{x}) + b(\boldsymbol{y})$ と抑えられる場合、$C[\pi]$ の最小値を与えるカップリングが存在する。$c(\boldsymbol{x},\boldsymbol{y}) = |\boldsymbol{x} - \boldsymbol{y}|^2$ ならば、$a(\boldsymbol{x}) = b(\boldsymbol{y}) = 0$ とすればよい。

1.1.1　距離であること

r-Wasserstein 距離は、距離と名乗る以上式 (5) に示した条件を満たす必要がある。

まず初めに非負性が成り立たなくてはならない。これは $d(\boldsymbol{x},\boldsymbol{y})$ および $\pi(\boldsymbol{x},\boldsymbol{y})$ の非負性から直ちに従う。

次に非退化性、すなわち「ゼロになるのは $p^{(0)}=p^{(1)}$ の場合に限る」という条件も必要となる。このことは、$d(\boldsymbol{x},\boldsymbol{y})=0 \iff \boldsymbol{x}=\boldsymbol{y}$ から、Wasserstein 距離がゼロならば $\pi(\boldsymbol{x},\boldsymbol{y})\neq 0 \iff \boldsymbol{x}=\boldsymbol{y}$ ということであり、これは輸送が行われないことを意味する。したがって $p^{(0)}=p^{(1)}$ が従う。

対称性 $\mathcal{W}_r(p^{(0)},p^{(1)})=\mathcal{W}_r(p^{(1)},p^{(0)})$ も簡単に確かめられる。$\pi^\dagger(\boldsymbol{x},\boldsymbol{y}):=\pi(\boldsymbol{y},\boldsymbol{x})$ とすると、$\pi\in\Pi(p^{(0)},p^{(1)}) \iff \pi^\dagger\in\Pi(p^{(1)},p^{(0)})$ となり、$d(\boldsymbol{x},\boldsymbol{y})=d(\boldsymbol{y},\boldsymbol{x})$ であることから、$\mathcal{C}_r[\pi]=\mathcal{C}_r[\pi^\dagger]$ が成り立つ。ここから $\mathcal{W}_r(p^{(0)},p^{(1)})=\inf_{\pi\in\Pi(p^{(0)},p^{(1)})}\mathcal{C}_r[\pi]^{1/r}=\inf_{\pi\in\Pi(p^{(0)},p^{(1)})}\mathcal{C}_r[\pi^\dagger]^{1/r}=\inf_{\pi^\dagger\in\Pi(p^{(1)},p^{(0)})}\mathcal{C}_r[\pi^\dagger]^{1/r}=\mathcal{W}_r(p^{(1)},p^{(0)})$ を得る。

最後に三角不等式であるが、これは次のように示される。そもそも三角不等式はこの場合、任意の $p^{(2)}$ に対して次が常に成り立つことを意味する：

$$\mathcal{W}_r(p^{(0)},p^{(1)})\leq\mathcal{W}_r(p^{(0)},p^{(2)})+\mathcal{W}_r(p^{(2)},p^{(1)}). \tag{8}$$

これを示すために、$\mathcal{W}_r(p^{(0)},p^{(2)})$ および $\mathcal{W}_r(p^{(2)},p^{(1)})$ にそれぞれ対応する最適カップリングとして π_0,π_1 を取る。このとき、$X\times X\times X$ 上の確率分布 γ であって $\int\gamma(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})d\boldsymbol{z}=\pi_0(\boldsymbol{x},\boldsymbol{y})$ および $\int\gamma(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})d\boldsymbol{x}=\pi_1(\boldsymbol{y},\boldsymbol{z})$ を満たすものが存在する。これは数学的には gluing lemma として知られる[3]が、素朴に考えれば $\gamma(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})=\pi_0(\boldsymbol{x},\boldsymbol{y})\pi_1(\boldsymbol{y},\boldsymbol{z})/p^{(2)}(\boldsymbol{y})$ とすればよい。さてこのとき、$\tilde\pi(\boldsymbol{x},\boldsymbol{y})=\int\gamma(\boldsymbol{x},\boldsymbol{z},\boldsymbol{y})d\boldsymbol{z}$ とすれば $\tilde\pi\in\Pi(p^{(0)},p^{(1)})$ となるので、

$$\begin{aligned}
\mathcal{W}_r(p^{(0)},p^{(1)}) &\leq \left[\int d(\boldsymbol{x},\boldsymbol{y})^r\tilde\pi(\boldsymbol{x},\boldsymbol{y})d\boldsymbol{x}d\boldsymbol{y}\right]^{1/r}=\left[\int d(\boldsymbol{x},\boldsymbol{y})^r\gamma(\boldsymbol{x},\boldsymbol{z},\boldsymbol{y})d\boldsymbol{x}d\boldsymbol{y}d\boldsymbol{z}\right]^{1/r}\\
&\leq\left[\int[d(\boldsymbol{x},\boldsymbol{z})+d(\boldsymbol{z},\boldsymbol{y})]^r\gamma(\boldsymbol{x},\boldsymbol{z},\boldsymbol{y})d\boldsymbol{x}d\boldsymbol{y}d\boldsymbol{z}\right]^{1/r}\\
&\leq\left[\int d(\boldsymbol{x},\boldsymbol{z})^r\gamma(\boldsymbol{x},\boldsymbol{z},\boldsymbol{y})d\boldsymbol{x}d\boldsymbol{y}d\boldsymbol{z}\right]^{1/r}+\left[\int d(\boldsymbol{z},\boldsymbol{y})^r\gamma(\boldsymbol{x},\boldsymbol{z},\boldsymbol{y})d\boldsymbol{x}d\boldsymbol{y}d\boldsymbol{z}\right]^{1/r}\\
&=\left[\int d(\boldsymbol{x},\boldsymbol{z})^r\pi_0(\boldsymbol{x},\boldsymbol{z})d\boldsymbol{x}d\boldsymbol{z}\right]^{1/r}+\left[\int d(\boldsymbol{z},\boldsymbol{y})^r\pi_1(\boldsymbol{z},\boldsymbol{y})d\boldsymbol{z}d\boldsymbol{y}\right]^{1/r}=\mathcal{W}_r(p^{(0)},p^{(2)})+\mathcal{W}_r(p^{(2)},p^{(1)})
\end{aligned}$$

より三角不等式が従う。ただし、2 行目では $d(\boldsymbol{x},\boldsymbol{y})$ の三角不等式を用いた。また、3 行目では Minkowski の不等式[4]を用いた。

1.1.2　そのほかの性質

$r\leq s$ のとき $\mathcal{W}_r(p^{(0)},p^{(1)})\leq\mathcal{W}_s(p^{(0)},p^{(1)})$ が成り立つ。このことは Hölder の不等式[5]を認めれば一瞬で示すことができる。$t=s/r$ とすると、$t\geq 1$ で、

$$\begin{aligned}
\left[\int d(\boldsymbol{x},\boldsymbol{y})^s\pi(\boldsymbol{x},\boldsymbol{y})d\boldsymbol{x}d\boldsymbol{y}\right]^{1/s}&=\left[\left[\int[d(\boldsymbol{x},\boldsymbol{y})^r]^t\pi(\boldsymbol{x},\boldsymbol{y})d\boldsymbol{x}d\boldsymbol{y}\right]^{1/t}\right]^{1/r}=\left[\|d^r\|_{L^t(\pi)}\right]^{1/r}\\
&\geq\left[\frac{\|d^r\|_{L^1(\pi)}}{\|1\|_{L^{1/(1-1/t)}(\pi)}}\right]^{1/r}=\left[\int d(\boldsymbol{x},\boldsymbol{y})^r\pi(\boldsymbol{x},\boldsymbol{y})d\boldsymbol{x}d\boldsymbol{y}\right]^{1/r}
\end{aligned}$$

が成り立つ。ただし二行目を得るのに Hölder の不等式 $\|d^r\|_{L^t(\pi)}\|1\|_{L^{1/(1-1/t)}(\pi)}\geq\|d^r\|_{L^1(\pi)}$ を用いた。

---

[3] [2, p.11]

[4] r-ノルムに関する三角不等式に他ならない。x を一般的な変数、$p(x)\geq 0$ とし、$\|f\|_{L^r(p)}=[\int|f(x)|^r p(x)dx]^{1/r}$ としたとき、$\|f+g\|_{L^r(p)}\leq\|f\|_{L^r(p)}+\|g\|_{L^r(p)}$. $x=(\boldsymbol{x},\boldsymbol{z},\boldsymbol{y})$, $p=\gamma$, $f(x)=d(\boldsymbol{x},\boldsymbol{z})$ および $g(x)=d(\boldsymbol{z},\boldsymbol{y})$ とすればよい。

[5] $r,s\in[1,\infty]$ が $r^{-1}+s^{-1}=1$ を満たす時、$\|f\|_{L^r(p)}\|g\|_{L^s(p)}\geq\|fg\|_{L^1(p)}$. Cauchy–Schwarz 不等式の一般化である（$r=s=1/2$ の場合そうなる）。

1.1.3 Monge 問題

Wasserstein 距離のより原始的な定式化として、次の **Monge 問題**がある。すなわち、場所 $\boldsymbol{x}$ からはかならず場所 $\boldsymbol{T}(\boldsymbol{x})$ に移すことにし、

$$\mathcal{M}[\boldsymbol{T}] = \int c(\boldsymbol{x}, \boldsymbol{T}(\boldsymbol{x}))p^{(0)}(\boldsymbol{x})d\boldsymbol{x} \tag{9}$$

によって総コストを与える。このとき、動かした後の分布が $p^{(1)}$ になるような $\boldsymbol{T}$ について $\mathcal{M}[\boldsymbol{T}]$ を最小化する。

ここで、「動かした後の分布が $p^{(1)}$ になる」ことは、$\boldsymbol{T}$ が単射（各 $\boldsymbol{y}$ に対し $\boldsymbol{y} = \boldsymbol{T}(\boldsymbol{x})$ となる $\boldsymbol{x}$ は高々一つ）の場合、Jacobian $J_{\boldsymbol{T}} = |\det(\nabla \boldsymbol{T})|$ を用いて、

$$p^{(1)}(\boldsymbol{y}) = \frac{p^{(0)}(\boldsymbol{T}^{-1}(\boldsymbol{y}))}{J_{\boldsymbol{T}}(\boldsymbol{T}^{-1}(\boldsymbol{y}))} \tag{10}$$

と表現できる。一般の場合も

$$p^{(1)}(\boldsymbol{y}) = \int p^{(0)}(\boldsymbol{x})\delta(\boldsymbol{y} - \boldsymbol{T}(\boldsymbol{x}))d\boldsymbol{x} \tag{11}$$

と書くことができるが、別の方法として、任意のテスト関数 f に対して

$$\int f(\boldsymbol{y})p^{(1)}(\boldsymbol{y})dy = \int f(\boldsymbol{T}(\boldsymbol{x}))p^{(0)}(\boldsymbol{x})d\boldsymbol{x} \tag{12}$$

が成り立つことを要請しても良い[6]。このような関係を以下では $p^{(1)} = \boldsymbol{T}\#p^{(0)}$ と表すことにする[7]。$\boldsymbol{T}$ がこの条件を満たす時、常に対応するカップリングが一つ存在し $\pi = (\mathrm{id} \times \boldsymbol{T})\#p^{(0)}$ によって与えられる[8]。

Monge 問題は、Dirac 関数的な分布を考えると破綻してしまう。たとえば、適当な $\boldsymbol{a} \in X\backslash\{\boldsymbol{0}\}$ に対し、$p^{(0)}(\boldsymbol{x}) = \delta(\boldsymbol{x})$, $p^{(1)}(\boldsymbol{x}) = [\delta(\boldsymbol{x} - \boldsymbol{a}) + \delta(\boldsymbol{x} + \boldsymbol{a})]/2$ とすると、$\pi(\boldsymbol{x}, \boldsymbol{y}) = \delta(\boldsymbol{x})[\delta(\boldsymbol{y} - \boldsymbol{a}) + \delta(\boldsymbol{y} + \boldsymbol{a})]/2$ が唯一のカップリングになる一方、$p^{(0)}$ を $p^{(1)}$ に移す $\boldsymbol{T}$ は存在しない。

一方、$c(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{x}, \boldsymbol{y})^r$ $(r > 1)$ の場合、$p^{(0)}, p^{(1)}$ がデルタ関数的な特異性を持たなければ Monge 問題と元々の最小化問題の解は一致し、一意的であることが知られている。$r = 1$ の場合は解に一意性がなく、一致する保証はない[9]。

1.2 Kantorovich 双対性

Wasserstein 距離の全く異なる表現である Kantorovich 双対性公式を説明する。

一般に最小化問題は別の最大化問題と結び付けられる。そのような関係を双対性（duality）と呼び、対をなす最適化問題を双対問題と呼ぶ。Wasserstein 距離は式 (6) において輸送コストの最小値として定義されたが、最大化問題としての定式化も可能であり、その双対性は **Kantorovich 双対性**と呼ばれる。すなわち一般に次が成り立つ：

$$\inf_{\pi \in \Pi(p^{(0)}, p^{(1)})} \int c(\boldsymbol{x}, \boldsymbol{y})\pi(\boldsymbol{x}, \boldsymbol{y})d\boldsymbol{x}d\boldsymbol{y} = \sup_{\phi, \psi} \left[\int \phi(\boldsymbol{y})p^{(1)}(\boldsymbol{y})dy - \int \psi(\boldsymbol{x})p^{(0)}(\boldsymbol{x})d\boldsymbol{x} \right] \quad \text{s.t. } \phi(\boldsymbol{y}) - \psi(\boldsymbol{x}) \leq c(\boldsymbol{x}, \boldsymbol{y}). \tag{13}$$

慣例に倣い、証明の前に感覚の説明を行う（図 2）。確率分布 p を砂山でなく、商品の量と考えよう（あるいは砂を商品と思っても良い）。式 (13) において左辺は相変わらず輸送、あるいは運送の総コストと解釈できる。輸送費の下限である。一方右辺では、まず $\psi(\boldsymbol{x})$ が土地 $\boldsymbol{x}$ における商品の仕入れ値と解釈される。ここでの仕入れ量は $p^{(0)}(\boldsymbol{x})$ である。そして、仕入れた商品をどうにか輸送し、土地 y において $p^{(1)}(\boldsymbol{y})$ の分量だけ、$\phi(\boldsymbol{y})$ の値段で売る。したがって右辺の目的関数は売買による利益に相当する。

---

[6] 任意のテスト関数に対して成り立つことを指して「弱い意味で weakly」と言ったりするが、確率分布とは確率変数の様々な関数に対して期待値を返すだけの道具と思えば、決して弱すぎることはない。もっとも、テスト関数の範囲の選び方には任意性があるだろう。

[7] # は押し出し（push forward）と呼ばれる。まさに $p^{(0)}$ を $p^{(1)}$ に運送するイメージである。

[8] つまり $\pi(\boldsymbol{x}, \boldsymbol{y}) = p^{(0)}(\boldsymbol{x})\delta(\boldsymbol{y} - \boldsymbol{T}(\boldsymbol{x}))$ である。

[9] この辺りの結果と証明は [2] の 9 章と 10 章にまとまっており、[3] の 6,7 ページにも要約が書かれている。本質的には、後で登場する c-凸関数の c-変換を与える点（c-劣微分）が一意であることが重要であり、このとき c-微分から最適輸送写像が得られる。$r = 2$ の場合に関する議論は 1.3.3 章および 1.3.4 章参照。

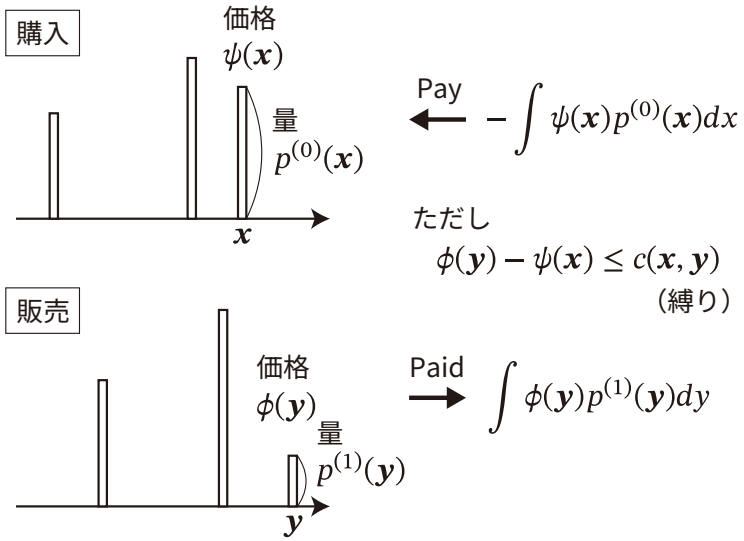

図 2　Kantorovich 双対性のイメージ図。場所 x では $\psi(x)$ の値段で $p^{(0)}(x)$ だけ仕入れる。それを場所 y に $p^{(1)}(y)$ だけ運び、価格 $\phi(y)$ で売る。このとき、運ぶコストは明示的には含まれていないが、価格設定において $\phi(y)-\psi(x)\le c(x,y)$ のセルフ縛りを課す。

売値 ϕ を高くつけたり、買値 ψ を値切ったりすれば利益が大きくできるわけだが、ここでは節度を持って商売を行うこととし、その限界を $\phi(y)-\psi(x)\le c(x,y)$ によって定める。つまり、仕入れ値 $\psi(x)$ に対し、売値を $\psi(x)+c(x,y)$ より高くはしないというルールを作る。こうすると、y へは複数の場所 x から商品が運ばれてくるかもしれないから、いつでも $\phi(y)=\psi(x)+c(x,y)$ を達成するのは難しそうである。したがって、このルールでは右辺のように売買による利益を最大化しても、左辺の輸送費を賄い切れるとは限らない。しかしながら、実際にこの等式が成り立つ。つまりは、輸送費の分しか上乗せしない値付けをしても（仕入れ値をも自由に操れるという仮定のもとでは）輸送費は必ず賄えるということを意味している[10]。

1.2.1　双対性の証明の概要

式 (13) の右辺の形はいかにも藪から棒に見えるが、ϕ と ψ は本質的には Lagrange の未定乗数にすぎない。カップリングの条件を陽に含んだ最適化の問題は、次の汎函数の最小化として表現される：

$$\mathcal{L}[\pi,\phi,\psi]=\mathcal{C}[\pi]+\int\phi(y)\left[p^{(1)}(y)-\int\pi(x,y)dx\right]dy-\int\psi(x)\left[p^{(0)}(x)-\int\pi(x,y)dy\right]dx. \tag{14}$$

最小化する上では $\pi\ge 0$ のみが条件として課される。この式は

$$\mathcal{L}[\pi,\phi,\psi]=\int[c(x,y)-\phi(y)+\psi(x)]\pi(x,y)dxdy+\int\phi(y)p^{(1)}(y)dy-\int\psi(x)p^{(0)}(x)dx \tag{15}$$

と改められるから、$c(x,y)-\phi(y)+\psi(x)<0$ であれば π をその点に集中させることでいくらでも小さくできる。元の $\inf_{\pi\in\Pi(p^{(0)},p^{(1)})}\mathcal{C}[\pi]$ を再現するには、ϕ,ψ に関する最大化を行い、少なくとも $c(x,y)-\phi(y)+\psi(x)\ge 0$ という条件が満たされなくてはならない。実際には $\inf_{\pi\ge 0}\sup_{\phi,\psi}\mathcal{L}=\inf_{\pi\in\Pi(p^{(0)},p^{(1)})}\mathcal{C}[\pi]$ が成り立つのだが、ここでさらに最小化と最大化の順番を入れ替えられるとすると、$c(x,y)-\phi(y)+\psi(x)>0$ の点において $\pi(x,y)=0$ とすれば π に関する最小化は完了する。あとは ϕ,ψ に関する最大化を $c(x,y)-\phi(y)+\psi(x)\ge 0$ という条件のもと行えば、元々の $\inf_{\pi\in\Pi(p^{(0)},p^{(1)})}\mathcal{C}[\pi]$ が復元されると期待される。その結果が式 (13) である。

実際の数学的証明はより繊細な注意を要する。上記の議論をフォーマルにすることで、片側の不等式（式 (13) の左辺≥右辺）は示されるものの、逆向きの証明は結構大変である。前者の結果は（最小化問題の結果）≥（最大化問題の結果）という不等号に対応するが、この時のギャップは双対性ギャップと呼ばれ、ギャップが閉じる時、強双対性

[10] そのような値付けをしなければ、人々は自分で買って自分で輸送してしまう、という説明もあるが、個人的には善良な商売人は損をしないというストーリーの方がしっくりくる。

が成り立つという。Kantorovich 双対の強双対性の一般的な証明は [2, Thm.5.10] を参照。離散の場合は通常の線形計画問題になるため、その一般論が適用できる[11]。

1.2.2 相補性

上述のおおざっぱな議論から、最適解 π^*, ϕ^*, ψ^* に関して、$c(\boldsymbol{x},\boldsymbol{y}) - \phi^*(\boldsymbol{y}) + \psi^*(\boldsymbol{x}) > 0$ が成り立つ点では $\pi^*(\boldsymbol{x},\boldsymbol{y}) = 0$ となることが予測された。$c(\boldsymbol{x},\boldsymbol{y}) - \phi^*(\boldsymbol{y}) + \psi^*(\boldsymbol{x}) \geq 0$ であることから、$\pi^*(\boldsymbol{x},\boldsymbol{y}) \neq 0$ である場合 $c(\boldsymbol{x},\boldsymbol{y}) - \phi^*(\boldsymbol{y}) + \psi^*(\boldsymbol{x}) = 0$ でなくてはならない[12]。このような関係性は相補性（complimentary slackness）と呼ばれ、実際に成り立つことが知られている [2, Thm.5.10]。

これら二つの条件は、ある集合 $\Gamma \subset X \times X$ によって特徴づけられることも知られている。すなわち、ある Γ が存在して、その中では $\phi^*(\boldsymbol{y}) - \psi^*(\boldsymbol{x}) = c(\boldsymbol{x},\boldsymbol{y})$ が成立し、外では $\pi^*(\boldsymbol{x},\boldsymbol{y}) = 0$ が成立する。さらに、この集合 Γ は c-循環的単調集合として特徴づけられる。$\Gamma \subset X \times X$ が c-循環的単調集合であるとは、任意の $\{(\boldsymbol{x}_i, \boldsymbol{y}_i) \in X \times X\}_{i=1}^k$ に対し、

$$\sum_i c(\boldsymbol{x}_i, \boldsymbol{y}_i) \leq \sum_i c(\boldsymbol{x}_i, \boldsymbol{y}_{i+1}) \tag{16}$$

が成り立つことを言う。これはざっくり言ってしまえば、2 点のペアがたくさん入った集合 Γ において、いずれのペアも最適な輸送に対応しており、これらを乱してしまえばコストは上昇する、ということである。

1.2.3 関数を減らす

ここまで考えていた双対問題には、ϕ と ψ という二つの関数が出てきたが、以下のように一つの関数のみによる最適化問題とすることもできる：

$$\inf_{\pi \in \Pi(p^{(0)}, p^{(1)})} \int c(\boldsymbol{x},\boldsymbol{y})\pi(\boldsymbol{x},\boldsymbol{y})d\boldsymbol{x}d\boldsymbol{y} = \sup_{\psi} \left[\int \psi^c(\boldsymbol{y}) p^{(1)}(\boldsymbol{y}) d\boldsymbol{y} - \int \psi(\boldsymbol{x}) p^{(0)}(\boldsymbol{x}) d\boldsymbol{x} \right]. \tag{17}$$

ただしここで ψ^c は ψ の c-変換と呼ばれる関数で、以下のように定義される：

$$\psi^c(\boldsymbol{y}) \coloneqq \inf_{\boldsymbol{x}} [\psi(\boldsymbol{x}) + c(\boldsymbol{x},\boldsymbol{y})]. \tag{18}$$

$c(\boldsymbol{x},\boldsymbol{y}) = |\boldsymbol{x} - \boldsymbol{y}|^2$ の場合、これは

$$\psi^c(\boldsymbol{y}) = |\boldsymbol{y}|^2 + \inf_{\boldsymbol{x}} \left[-2\boldsymbol{x} \cdot \boldsymbol{y} + |\boldsymbol{x}|^2 + \psi(\boldsymbol{x}) \right] \tag{19}$$

となり、$[|\boldsymbol{x}|^2 + \psi(\boldsymbol{x})]/2$ という関数の Legendre 変換と似ていることがわかる。

この関係を理解するためにはまず、$p^{(1)}(\boldsymbol{y}) > 0$ であるとき、$\pi^*(\boldsymbol{x},\boldsymbol{y})$ がどこかの $\boldsymbol{x} = \boldsymbol{x}^*$ ではゼロでないことに注意する。すると相補性から $\boldsymbol{x}^*$ では $\phi^*(\boldsymbol{y}) - \psi^*(\boldsymbol{x}^*) = c(\boldsymbol{x}^*, \boldsymbol{y})$ が成り立つことがわかる。一般に $\phi^*(\boldsymbol{y}) - \psi^*(\boldsymbol{x}) \leq c(\boldsymbol{x},\boldsymbol{y})$ だったことを思い出すと、$\phi^*(\boldsymbol{y}) = \inf_{\boldsymbol{x}}[\psi^*(\boldsymbol{x}) + c(\boldsymbol{x},\boldsymbol{y})]$ と書ける。$p^{(1)}(\boldsymbol{y}) = 0$ である場合は関係がないので、任意の $\boldsymbol{y}$ において $\phi^* = [\psi^*]^c$ が成り立つと考えても問題がない。すなわち最大化において、ϕ と ψ を別個に探す必要はなく、選んだ ψ に対して ψ^c を ϕ の位置に入れた値を最大化すれば十分であることがわかる。

さらに、この考察を $p^{(0)}(\boldsymbol{x}) > 0$ の場合に行うことで、$\psi^*(\boldsymbol{x}) = \sup_{\boldsymbol{y}}[\phi^*(\boldsymbol{y}) - c(\boldsymbol{x},\boldsymbol{y})]$ となることがわかる。一般に関数 ψ が c-凸であるとは、ある関数 ζ が存在して

$$\psi(\boldsymbol{x}) = \sup_{\boldsymbol{y}} [\zeta(\boldsymbol{y}) - c(\boldsymbol{x},\boldsymbol{y})] \tag{20}$$

と書けることを言う。つまり最適解 ψ^* は c-凸であり、最大化 (17) においても ψ は c-凸であるという制限をおいても良い。

---

[11] 弱双対性までは [1] にある。強双対性については、たとえば [4] を参照。

[12] このように、各点で等号が成り立つ、というような主張は実は危ない。空間に穴がプツリと開いていても、本来確率分布（より正確には測度）にとって重要でないから、そのことを留意した記述が求められる。相補性であれば、$c(\boldsymbol{x},\boldsymbol{y}) - \phi^*(\boldsymbol{y}) + \psi^*(\boldsymbol{x}) \neq 0$ となる点を全部集めて Z という集合を作っても、$\int_Z \pi(\boldsymbol{x},\boldsymbol{y})d\boldsymbol{x}d\boldsymbol{y} = 0$ である、と言った方が正確である。

1.2.4 c-凸条件, $c = d$ のとき

$c(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{x}, \boldsymbol{y})$ の場合、c-凸という性質は 1-Lipschitz であることと等価になる。一般に関数 ψ が K-Lipschitz であるとは、

$$\|\psi\|_{\mathrm{Lip}} \coloneqq \sup_{\boldsymbol{x},\boldsymbol{y}} \frac{|\psi(\boldsymbol{x}) - \psi(\boldsymbol{y})|}{d(\boldsymbol{x}, \boldsymbol{y})} \leq K \tag{21}$$

が成り立つことを指す。$\|\cdot\|_{\mathrm{Lip}}$ を Lipschitz ノルムと呼び、ψ が微分可能で、d が Euclid 距離の場合 $\|\psi\|_{\mathrm{Lip}} = \sup_{\boldsymbol{x}} |\nabla\psi(\boldsymbol{x})|$ が成り立つ。また、c-変換は自分自身になる。つまり $c = d$ ならば $\psi^c = \psi$ である。証明はやや技巧的だが、あまり難しくはない。

まず、c-凸性の定義から 1-Lipschitz 性を導く。ψ が c-凸であるとし、適当な点の組 $\boldsymbol{x}, \boldsymbol{y}$ に関して $\psi(\boldsymbol{x}) \geq \psi(\boldsymbol{y})$ が成り立つとする。c-凸性の定義において最大値を与える点が存在し、$\psi(\boldsymbol{x}) = \sup_{\boldsymbol{y}} [\zeta(\boldsymbol{y}) - d(\boldsymbol{x}, \boldsymbol{y})] = \zeta(\boldsymbol{x}^*) - d(\boldsymbol{x}, \boldsymbol{x}^*)$ と書けるとすると、

$$\psi(\boldsymbol{x}) - \psi(\boldsymbol{y}) = \zeta(\boldsymbol{x}^*) - d(\boldsymbol{x}, \boldsymbol{x}^*) - \psi(\boldsymbol{y}) \leq \zeta(\boldsymbol{x}^*) - d(\boldsymbol{x}, \boldsymbol{x}^*) - (\zeta(\boldsymbol{x}^*) - d(\boldsymbol{y}, \boldsymbol{x}^*)) = -d(\boldsymbol{x}, \boldsymbol{x}^*) + d(\boldsymbol{y}, \boldsymbol{x}^*) \tag{22}$$

が成り立つ。ただし不等式は c-凸性の定義 (20) から従う。ここに三角不等式 $d(\boldsymbol{y}, \boldsymbol{x}^*) \leq d(\boldsymbol{x}, \boldsymbol{y}) + d(\boldsymbol{x}, \boldsymbol{x}^*)$ を適用すれば、$|\psi(\boldsymbol{x}) - \psi(\boldsymbol{y})| \leq d(\boldsymbol{x}, \boldsymbol{y})$ が従う。$\psi(\boldsymbol{x}) < \psi(\boldsymbol{y})$ の場合も同様に示せるため、c-凸ならば 1-Lipschitz である。逆に、1-Lipschitz である場合、$|\psi(\boldsymbol{x}) - \psi(\boldsymbol{y})| \leq d(\boldsymbol{x}, \boldsymbol{y})$ より、$\psi(\boldsymbol{x}) \geq \psi(\boldsymbol{y}) - d(\boldsymbol{x}, \boldsymbol{y})$ が従う[13]。よって、$\psi(\boldsymbol{x}) = \sup_{\boldsymbol{y}}[\psi(\boldsymbol{y}) - d(\boldsymbol{x}, \boldsymbol{y})]$ が成り立つ（$\boldsymbol{y} = \boldsymbol{x}$ において最大値が与えられる）。したがって ψ が c-凸である。また不等式 $\psi(\boldsymbol{x}) \geq \psi(\boldsymbol{y}) - d(\boldsymbol{x}, \boldsymbol{y})$ は $\psi(\boldsymbol{y}) \leq \psi(\boldsymbol{x}) + d(\boldsymbol{x}, \boldsymbol{y})$ と書けるので、$\psi(\boldsymbol{y}) = \inf_{\boldsymbol{x}}[\psi(\boldsymbol{x}) + d(\boldsymbol{x}, \boldsymbol{y})]$ となり、1-Lipschitz な関数の c-変換は自分自身であることもわかる。

1.2.5 期待値による表現

以上の結果から、次の関係を得る：

$$\mathcal{W}_1(p^{(0)}, p^{(1)}) = \sup_{\psi:\|\psi\|_{\mathrm{Lip}} \leq 1} \int \psi(\boldsymbol{x})[p^{(1)}(\boldsymbol{x}) - p^{(0)}(\boldsymbol{x})]dx. \tag{23}$$

最大化は単一の関数で十分であり、その関数は c-凸であること、そして $c = d$ の時 c-凸性は 1-Lipschitz 性に帰着し、c-変換は 1-Lipschitz 関数に対して恒等変換であることを考えれば直ちに導かれる。この関係性を特に Kantorovich–Rubinstein 双対性と呼ぶ。

$p^{(i)}$ による期待値を $\langle\cdot\rangle_i$ と書くことにすると、この関係式は

$$\mathcal{W}_1(p^{(0)}, p^{(1)}) = \sup_{\psi:\|\psi\|_{\mathrm{Lip}} \leq 1} [\langle\psi\rangle_1 - \langle\psi\rangle_0] \tag{24}$$

と書ける。Lipschitz ノルムの性質から、次のようにも表すこともできる：

$$\mathcal{W}_1(p^{(0)}, p^{(1)}) = \sup_{\psi} \frac{\langle\psi\rangle_1 - \langle\psi\rangle_0}{\|\psi\|_{\mathrm{Lip}}}. \tag{25}$$

すなわち、単一の物理量 ψ に着目し、その期待値の変化と傾きの最大値の情報を得ることで、二状態間の 1-Wasserstein 距離を推定することができる。

この関係式は、d が距離関数であれば常に成り立ち、d は Euclid 距離でなくてもよい。特に、X は離散的な空間でもよく、のちに d としてグラフ上の距離を考える場合もこれまでの議論は含んでいる。

1.3 運動論的な定式化

ここまで、輸送においては始まりと終わりの状態にだけ注目してきた。一方、その間の「時間発展」を考えることによっても Wasserstein 距離は定式化が可能である。距離関数を Euclid 距離 $d(\boldsymbol{x}, \boldsymbol{y}) = |\boldsymbol{x} - \boldsymbol{y}| = \sqrt{\sum_{i=1}^N |x_i - y_i|^2}$

[13] 数直線上に $\psi(\boldsymbol{x})$ と $\psi(\boldsymbol{y})$ を置くと、これらは $d(\boldsymbol{x}, \boldsymbol{y})$ も離れていない。したがって $\psi(\boldsymbol{y})$ を $d(\boldsymbol{x}, \boldsymbol{y})$ だけ後退させれば必ず $\psi(\boldsymbol{x})$ の後ろに行く。

とし、滑らかな分布を考えた場合、r-Wasserstein 距離は次のように与えることができる：

$$\mathcal{W}_r(p^{(0)}, p^{(1)}) = \inf_{\{p_t\},\{\boldsymbol{v}_t\}} \left[\int_0^1 \int p_t(\boldsymbol{x}) |\boldsymbol{v}_t(\boldsymbol{x})|^r dx dt \right]^{1/r}, \tag{26}$$

ただし、$t \in [0,1]$ において定義される確率分布 p_t およびベクトル場 $\boldsymbol{v}_t$ の時間発展は、境界条件 $p_0 = p^{(0)}$, $p_1 = p^{(1)}$ および次の連続の式を満たす：

$$\partial_t p_t(\boldsymbol{x}) = -\nabla \cdot (p_t(\boldsymbol{x}) \boldsymbol{v}_t(\boldsymbol{x})). \tag{27}$$

特に $r = 2$ を指して Benamou–Brenier 公式と呼ぶことが多い。

連続の式 (27) は密度分布 $p_t(\boldsymbol{x})$ が速度場 $\boldsymbol{v}_t(\boldsymbol{x})$ によって運ばれる状況に対応する。ここでは詳細の説明は行わないが、p_t を確率分布とみなした場合、連続の式は Fokker–Planck 方程式と呼ばれる、確率的時間発展の表現に相当する。$r = 2$ の場合、コスト関数 $\int p_t(\boldsymbol{x}) |\boldsymbol{v}_t(\boldsymbol{x})|^2 dx$ はブラウン運動のような熱的な確率ダイナミクスにおけるエントロピーの単位時間あたりの増大（の定数倍）を与える。すなわち、2-Wasserstein 距離は、確率分布を変化させる際の最小自由エネルギー散逸を与える（詳細は [5] など参照）。

1.3.1 時間幅に関する不定性

この公式は時間幅に関する不定性が存在する。$t' = \tau t$ としたとき、連続の式を満たす $p_t, \boldsymbol{v}_t$ を $p'_{t'} = p_{t'/\tau}$, $v'_{t'} = v_{t'/\tau}/\tau$ と置き換えれば $p'_{t'}, v'_{t'}$ もまた連続の式を満たす。したがって、式 (26) は次のように書き表すことができる：

$$\mathcal{W}_r(p^{(0)}, p^{(1)}) = \inf_{\{p_t\},\{\boldsymbol{v}_t\}} \left[\tau^{r-1} \int_0^\tau \int p_t(\boldsymbol{x}) |\boldsymbol{v}_t(\boldsymbol{x})|^r dx dt \right]^{1/r}. \tag{28}$$

1.3.2 ポテンシャルによる表現

連続の式では、速度場がポテンシャルの勾配によって代替可能であることが重要である（ただしある程度性質の良い境界条件は必要で、ここでは $|\boldsymbol{x}| \to 0$ で $p_t(\boldsymbol{x}) \to 0$ を仮定する）。関数 ϕ に対する偏微分方程式

$$\nabla \cdot (p_t(\boldsymbol{x}) \nabla \phi(\boldsymbol{x})) = -\partial_t p_t(\boldsymbol{x}) \tag{29}$$

を考えると、これはほとんど Poisson 方程式である。その解[14]を $\phi_t(\boldsymbol{x})$ とすれば、$\boldsymbol{v}_t$ は $\nabla \phi_t$ によって代替することができる。このことは $\boldsymbol{v}_t$ が $\nabla \phi_t$ という形で書けることを意味しない。むしろ、$\boldsymbol{v}_t = \nabla \phi_t + \boldsymbol{w}_t$ としたとき、$\boldsymbol{w}_t$ が

$$\nabla \cdot (p_t(\boldsymbol{x}) \boldsymbol{w}_t(\boldsymbol{x})) = 0 \tag{30}$$

を満たすことから、$\boldsymbol{v}_t$ は「無駄な」成分を含みうることを意味する。

この「無駄」さは $r = 2$ の場合実際に示すことができる。このとき

$$\begin{aligned} &\int p_t(\boldsymbol{x}) |\boldsymbol{v}_t(\boldsymbol{x})|^2 dx = \int p_t(\boldsymbol{x}) |\nabla \phi_t(\boldsymbol{x}) + \boldsymbol{w}_t(\boldsymbol{x})|^2 dx \\ &= \int p_t(\boldsymbol{x}) |\nabla \phi_t(\boldsymbol{x})|^2 dx + \int p_t(\boldsymbol{x}) |\boldsymbol{w}_t(\boldsymbol{x})|^2 dx + 2 \int p_t(\boldsymbol{x}) \nabla \phi_t(\boldsymbol{x}) \cdot \boldsymbol{w}_t(\boldsymbol{x}) dx \end{aligned}$$

が成り立つが、2 行目の第 3 項に対して部分積分を行うと $\boldsymbol{w}_t$ の性質より

$$\int p_t(\boldsymbol{x}) \nabla \phi_t(\boldsymbol{x}) \cdot \boldsymbol{w}_t(\boldsymbol{x}) dx = -\int \phi_t(\boldsymbol{x}) \nabla \cdot (p_t(\boldsymbol{x}) \boldsymbol{w}_t(\boldsymbol{x})) dx = 0 \tag{31}$$

を得る。したがって、

$$\int p_t(\boldsymbol{x}) |\boldsymbol{v}_t(\boldsymbol{x})|^2 dx = \int p_t(\boldsymbol{x}) |\nabla \phi_t(\boldsymbol{x})|^2 dx + \int p_t(\boldsymbol{x}) |\boldsymbol{w}_t(\boldsymbol{x})|^2 dx \geq \int p_t(\boldsymbol{x}) |\nabla \phi_t(\boldsymbol{x})|^2 dx \tag{32}$$

[14] 常識的に考えれば解は必ず存在しそうである。適切な可積分性を持った解の存在が言えれば一意性は Wikipedia “Uniqueness theorem for Poisson’s equation” と全く同様にして示せる。

が成り立ち、$\boldsymbol{w}_t$ がない方が全体の値を小さくできることがわかる。

以上の議論から、2-Wasserstein 距離は次のようにも与えられることがわかる：

$$\mathcal{W}_2(p^{(0)}, p^{(1)}) = \inf_{\{p_t\},\{\phi_t\}} \left[\int_0^1 \int p_t(\boldsymbol{x}) |\nabla \phi_t(\boldsymbol{x})|^2 d\boldsymbol{x} dt \right]^{1/2}, \tag{33}$$

ここで、$t \in [0,1]$ において定義される確率分布 p_t およびポテンシャル場 ϕ_t は、境界条件 $p_0 = p^{(0)}$, $p_1 = p^{(1)}$ および次の連続の式を満たさなくてはならない

$$\partial_t p_t(\boldsymbol{x}) = -\nabla \cdot (p_t(\boldsymbol{x}) \nabla \phi_t(\boldsymbol{x})). \tag{34}$$

1.3.3 c-凸性、Monge 問題、双対性

実は、今考えているダイナミクスの最適化は、Monge 問題や Kantorovich 双対性と直接的に関係する。そのことを明らかにするため、$c(\boldsymbol{x}, \boldsymbol{y}) = |\boldsymbol{x} - \boldsymbol{y}|^2$ の場合の c-凸性を再び考察してみよう。このとき c-凸性は通常の凸性と対応づけられ、双対性公式における最適な c-凸関数から凸関数が一つ構成される。そしてその勾配が Monge 問題の最適輸送写像になる。Benamou–Brenier 公式はそれらの概念から証明することができる。

さて、関数 ψ が c-凸であるとは、ある関数 ζ が存在して

$$\psi(\boldsymbol{x}) = \sup_{\boldsymbol{y}} [\zeta(\boldsymbol{y}) - c(\boldsymbol{x}, \boldsymbol{y})] \tag{35}$$

が成り立つことを言うのだった。ここに $c(\boldsymbol{x}, \boldsymbol{y}) = |\boldsymbol{x} - \boldsymbol{y}|^2$ を代入すると、

$$\psi(\boldsymbol{x}) = \sup_{\boldsymbol{y}} [\zeta(\boldsymbol{y}) - |\boldsymbol{x}|^2 - |\boldsymbol{y}|^2 + 2\boldsymbol{x} \cdot \boldsymbol{y}] \tag{36}$$

となる。いま、$\varphi(\boldsymbol{x}) = \frac{1}{2}[\psi(\boldsymbol{x}) + |\boldsymbol{x}|^2]$ とすると、この条件は φ がある関数 ξ を用いて

$$\varphi(\boldsymbol{x}) = \sup_{\boldsymbol{y}} [\boldsymbol{x} \cdot \boldsymbol{y} - \xi(\boldsymbol{y})] \tag{37}$$

と書けることと等価になる。ξ と ζ は $2\xi(\boldsymbol{y}) = -\zeta(\boldsymbol{y}) + |\boldsymbol{y}|^2$ によって結びつく。さてこのとき

$$\begin{aligned}
\varphi(\lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2) &= \sup_{\boldsymbol{y}} \{[\lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2] \cdot \boldsymbol{y} - \xi(\boldsymbol{y})\} \\
&= \sup_{\boldsymbol{y}} \{\lambda[\boldsymbol{x}_1 \cdot \boldsymbol{y} - \xi(\boldsymbol{y})] + (1-\lambda)[\boldsymbol{x}_2 \cdot \boldsymbol{y} - \xi(\boldsymbol{y})]\} \\
&\le \lambda \sup_{\boldsymbol{y}} [\boldsymbol{x}_1 \cdot \boldsymbol{y} - \xi(\boldsymbol{y})] + (1-\lambda) \sup_{\boldsymbol{y}} [\boldsymbol{x}_2 \cdot \boldsymbol{y} - \xi(\boldsymbol{y})] \\
&= \lambda \varphi(\boldsymbol{x}_1) + (1-\lambda)\varphi(\boldsymbol{x}_2)
\end{aligned}$$

が成り立つので、φ は通常の意味で凸関数になる。すなわち、c が Euclid 距離の 2 乗で与えられる時、Kantorovich に登場する c-凸関数 ψ は凸関数 φ と結びつく。また、c-変換 $\psi^c(\boldsymbol{y}) = \inf_{\boldsymbol{x}}[\psi(\boldsymbol{x}) + c(\boldsymbol{x}, \boldsymbol{y})]$ と、φ の Legendre 変換 $\varphi^{\mathrm{L}}(\boldsymbol{y}) = \sup_{\boldsymbol{x}}[\boldsymbol{x} \cdot \boldsymbol{y} - \varphi(\boldsymbol{x})]$ との間には関係式 $\varphi^{\mathrm{L}}(\boldsymbol{y}) = \frac{1}{2}[|\boldsymbol{y}|^2 - \psi^c(\boldsymbol{y})]$ が成り立つ。

ここからは各種の最適化問題の最適解に注目する。まず Kantorovich 双対性における相補性条件から、$\pi(\boldsymbol{x}, \boldsymbol{y}) \neq 0$ の場合（つまり $p^{(0)}$ および $p^{(1)}$ が 0 でない点のペアでは）$\psi^c(\boldsymbol{y}) - \psi(\boldsymbol{x}) = c(\boldsymbol{x}, \boldsymbol{y})$ が成り立つ。このことを対応する凸関数の言葉で書けば $\varphi^{\mathrm{L}}(\boldsymbol{y}) + \varphi(\boldsymbol{x}) = \boldsymbol{x} \cdot \boldsymbol{y}$ となる。Legendre 変換の一般論から、φ が狭義凸で十分滑らかであればこのような $\boldsymbol{y}$ は各 $\boldsymbol{x}$ に対してひとつしかなく、$\nabla\varphi(\boldsymbol{x}) = \boldsymbol{y}$ を満たす[15]ため、実は $\pi(\boldsymbol{x}, \boldsymbol{y}) \neq 0$ となりうる $\boldsymbol{y}$ は $\nabla\varphi(\boldsymbol{x})$ に絞られる。つまり、この対応関係が Monge 問題の最適輸送を与える。このことは次のように表現できる：

$$\pi = (\mathrm{id} \times \nabla\varphi)\# p^{(0)}, \quad p^{(1)} = \nabla\varphi\# p^{(0)}, \tag{38}$$

$$\nabla\varphi = \arg\min_{\boldsymbol{T}} \int |\boldsymbol{x} - \boldsymbol{T}(\boldsymbol{x})|^2 p^{(0)}(\boldsymbol{x}) d\boldsymbol{x} \ \text{ s.t. } \ p^{(1)} = \boldsymbol{T}\# p^{(0)}. \tag{39}$$

[15] このことは次のように理解できる。まず、$\varphi^{\mathrm{L}}(\boldsymbol{y}) = \sup_{\boldsymbol{x}}[\boldsymbol{x} \cdot \boldsymbol{y} - \varphi(\boldsymbol{x})]$ において、$\boldsymbol{x} \cdot \boldsymbol{y} - \varphi(\boldsymbol{x})$ は $\boldsymbol{x}$ の凹関数である。したがって、最大値を与える $\boldsymbol{x} = \boldsymbol{x}^*$ が存在する場合、そこが「頂点」になり微分は 0 になる。すなわち、$\boldsymbol{y} = \nabla\varphi(\boldsymbol{x}^*)$ が成り立つ。また、φ が狭義凸である場合、この「頂点」は一点であり、$\boldsymbol{y}$ と $\boldsymbol{x}^*$ の対応関係は一対一となる。このとき $\varphi(\boldsymbol{x}) + \varphi^{\mathrm{L}}(\boldsymbol{y}) = \boldsymbol{x}^* \cdot \boldsymbol{y}$ が成立し、このようなときにしか成立しない。この関係性は対称であり、$\boldsymbol{x}$ から $\boldsymbol{y}$ への対応は $\nabla\varphi^{\mathrm{L}}(\boldsymbol{y})$ によって与えられる。すなわち $(\nabla\varphi)^{-1} = \nabla\varphi^{\mathrm{L}}$ が成り立つ。

1.3.4 Benamou–Brenier 公式との関係

ここで得られた φ を用いて、Benamou–Brenier 公式の「証明」を行う。

$$\begin{aligned}&\inf_{\boldsymbol{T}} \int |\boldsymbol{x}-\boldsymbol{T}(\boldsymbol{x})|^2 p^{(0)}(\boldsymbol{x})d\boldsymbol{x} \quad \text{s.t.} \quad p^{(1)} = \boldsymbol{T}\#p^{(0)} \\ &= \inf_{\{p_t\},\{\boldsymbol{v}_t\}} \int_0^1 \int p_t(\boldsymbol{x})|\boldsymbol{v}_t(\boldsymbol{x})|^2 d\boldsymbol{x}dt \quad \text{s.t.} \quad \partial_t p_t(\boldsymbol{x}) = -\nabla\cdot(p_t(\boldsymbol{x})\boldsymbol{v}_t(\boldsymbol{x})),\ p_0 = p^{(0)},\ p_1 = p^{(1)}\end{aligned} \tag{40}$$

を示す。

まず、左辺≥右辺を示す。左辺の最適輸送写像は凸関数 φ によって $\nabla\varphi$ と与えられる。この φ を用いて $p_t, \boldsymbol{v}_t$ を以下のように定める。まず $\varphi_t = t\varphi + (1-t)\frac{|\boldsymbol{x}|^2}{2}$ とし、$\boldsymbol{\Phi}_t = \nabla\varphi_t$ とする。このとき φ_t は凸関数であり[16]、$\boldsymbol{\Phi}_t^{-1}$ は φ_t の Legendre 変換 φ_t^{L} の勾配として与えられる[17]。φ_t の形から、$\partial_t\boldsymbol{\Phi}_t = \nabla\varphi - \boldsymbol{x}$ が成り立つ。

さてこのとき、$p_t = \boldsymbol{\Phi}_t\#p^{(0)}$ および $\boldsymbol{v}_t(\boldsymbol{x}) = [\partial_t\boldsymbol{\Phi}_t](\boldsymbol{\Phi}_t^{-1}(\boldsymbol{x}))$ とすると、これらは連続の式を満たす[18]。すなわち、任意のテスト関数 f に対し

$$\int f(\boldsymbol{x})\partial_t p_t(\boldsymbol{x})d\boldsymbol{x} = -\int f(\boldsymbol{x})\nabla\cdot(p_t(\boldsymbol{x})\boldsymbol{v}_t(\boldsymbol{x}))d\boldsymbol{x} \tag{41}$$

が成り立つ。実際、# の定義[19]に注意すると、

$$\begin{aligned}\int f(\boldsymbol{x})\partial_t p_t(\boldsymbol{x})d\boldsymbol{x} &= \frac{d}{dt}\int f(\boldsymbol{\Phi}_t(\boldsymbol{x}))p^{(0)}(\boldsymbol{x})d\boldsymbol{x} = \int [\nabla f](\boldsymbol{\Phi}_t(\boldsymbol{x}))\cdot\partial_t\boldsymbol{\Phi}_t(\boldsymbol{x})p^{(0)}(\boldsymbol{x})d\boldsymbol{x} \\ &= \int [\nabla f](\boldsymbol{x})\cdot[\partial_t\boldsymbol{\Phi}_t](\boldsymbol{\Phi}_t^{-1}(\boldsymbol{x}))p_t(\boldsymbol{x})d\boldsymbol{x} = -\int f(\boldsymbol{x})\nabla\cdot(p_t(\boldsymbol{x})\boldsymbol{v}_t(\boldsymbol{x}))d\boldsymbol{x}\end{aligned}$$

となるので示される。初期条件は自明に満たされており、終条件も $\boldsymbol{\Phi}_1 = \nabla\varphi$ より満たされていることがわかる。

この $p_t, \boldsymbol{v}_t$ に関して目的関数を計算すると

$$\begin{aligned}&\int_0^1\int p_t(\boldsymbol{x})|\boldsymbol{v}_t(\boldsymbol{x})|^2 d\boldsymbol{x}dt = \int_0^1\int p_t(\boldsymbol{x})\left|[\partial_t\boldsymbol{\Phi}_t](\boldsymbol{\Phi}_t^{-1}(\boldsymbol{x}))\right|^2 d\boldsymbol{x}dt \\ &= \int_0^1\int p^{(0)}(\boldsymbol{x})\left|[\partial_t\boldsymbol{\Phi}_t](\boldsymbol{x})\right|^2 d\boldsymbol{x}dt = \int_0^1\int p^{(0)}(\boldsymbol{x})\left|\nabla\varphi(\boldsymbol{x})-\boldsymbol{x}\right|^2 d\boldsymbol{x}dt\end{aligned}$$

が成り立つから、この時点で右辺の最小化は左辺の値よりも大きくならないことが保障される。

次に、左辺≤右辺を示す。任意の連続の式を満たす組 $p_t, \boldsymbol{v}_t$ に対応する粒子の運動を考える。すなわち、$\boldsymbol{\phi}_t(\boldsymbol{x})$ を次の方程式の解とする：

$$\partial_t\boldsymbol{\phi}_t(\boldsymbol{x}) = \boldsymbol{v}_t(\boldsymbol{\phi}_t(\boldsymbol{x})), \quad \boldsymbol{\phi}_0(\boldsymbol{x}) = \boldsymbol{x}. \tag{42}$$

このとき、$p_t = \boldsymbol{\phi}_t\#p^{(0)}$ が成り立つ[20]。したがって、$\boldsymbol{T} = \boldsymbol{\phi}_1$ は Monge 問題の解の候補となる。$\boldsymbol{T} = \boldsymbol{\phi}_1$ とすると、

$$\begin{aligned}&\int |\boldsymbol{x}-\boldsymbol{T}(\boldsymbol{x})|^2 p^{(0)}(\boldsymbol{x})d\boldsymbol{x} = \int \left|\int_0^1 \boldsymbol{v}_t(\boldsymbol{\phi}_t(\boldsymbol{x}))dt\right|^2 p^{(0)}(\boldsymbol{x})d\boldsymbol{x} \\ &\le \int_0^1\int |\boldsymbol{v}_t(\boldsymbol{\phi}_t(\boldsymbol{x}))|^2 p^{(0)}(\boldsymbol{x})d\boldsymbol{x}dt = \int_0^1\int |\boldsymbol{v}_t(\boldsymbol{x})|^2 p_t(\boldsymbol{x})d\boldsymbol{x}dt\end{aligned}$$

が成り立つ。つまり、任意の連続の式を満たす組 $p_t, \boldsymbol{v}_t$ に対し、コストが大きくならない $\boldsymbol{T}$ を構成できるので、左辺≤右辺が示される。両向きの不等式が示されたので、等号が成り立つことがわかった。

---

[16] φ_t は Hessian が単位行列と φ の Hessian の凸結合になる。
[17] 脚注 15 参照
[18] 最適輸送写像による押し出し $((1-t)\boldsymbol{x} + t\boldsymbol{T}(\boldsymbol{x}))\#p^{(0)}$ を displacement interpolation と呼ぶ。
[19] 任意のテスト関数 f に対し $\int f(\boldsymbol{x})\boldsymbol{\Phi}_t\#p^{(0)}(\boldsymbol{x})d\boldsymbol{x} = \int f(\boldsymbol{\Phi}_t(\boldsymbol{x}))p^{(0)}(\boldsymbol{x})d\boldsymbol{x}$.
[20] これは偏微分方程式に対する特性曲線法と呼ばれる [3, Theorem 5.34]。

以上の証明から、Monge 問題の最適輸送写像 $\boldsymbol{T}$ と、Benamou–Brenier 公式の最適な速度場 $\boldsymbol{v}_t$ は、ともに Kantorovich 双対性における最適な c-凸関数 ψ（さらに言えばそこから構成される凸関数 φ）だけから構成することができ、本質的には同一の輸送を表現していることがわかる。この話題に関連した研究は広く深いが、語弊を恐れず言えば本質的にはここでやった話と似たようなものである。

問題

$p^{(0)}$ を平均 μ_0, 分散 σ_0^2 の一変数 Gauss 分布、$p^{(1)}$ を平均 μ_1, 分散 σ_1^2 の一変数 Gauss 分布としたとき、2-Wasserstein 距離は $\sqrt{(\mu_1-\mu_0)^2+(\sigma_1-\sigma_0)^2}$ によって与えられることが知られている [5]。このときもっとも自然な $\boldsymbol{T}$ を選ぶと Monge 問題がこの値を返すことを示し、対応する凸関数 φ およびダイナミクス $p_t, \boldsymbol{v}_t$ を求めよ。（ヒント：適当な座標変換によって $\mu_0=0, \sigma_0=1$ とできるから、そう仮定してよい。）

1.3.5 Hamilton–Jacobi 方程式

Benamou–Brenier 公式のポテンシャルによる表現 (33) における最適解が満たす方程式を求めよう。そのために、まず元々の Benamou–Brenier 公式 (26) に対して、次の Lagrange 未定乗数を用いた表現を考える：

$$\mathcal{W}_2(p^{(0)},p^{(1)})^2 = \inf_{\{p_t\},\{\boldsymbol{v}_t\}} \sup_{\{\phi_t\}} \mathcal{I}[p,\boldsymbol{v},\phi] \quad \text{with} \quad \mathcal{I}[p,\boldsymbol{v},\phi] = \int_0^1 \int \left[p_t|\boldsymbol{v}_t|^2 + 2\phi_t(\partial_t p_t + \nabla\cdot(p_t\boldsymbol{v}_t))\right] dxdt \tag{43}$$

p_t および $\boldsymbol{v}_t$ に関する極値条件は次のように書ける：

$$\frac{\delta \mathcal{I}}{\delta p_t(\boldsymbol{x})} = |\boldsymbol{v}_t(\boldsymbol{x})|^2 - 2\partial_t\phi_t(\boldsymbol{x}) - 2\boldsymbol{v}_t(\boldsymbol{x})\cdot\nabla\phi_t(\boldsymbol{x}) = 0, \tag{44}$$

$$\frac{\delta \mathcal{I}}{\delta \boldsymbol{v}_t(\boldsymbol{x})} = 2p_t(\boldsymbol{x})[\boldsymbol{v}_t(\boldsymbol{x}) - \nabla\phi_t(\boldsymbol{x})] = \boldsymbol{0}. \tag{45}$$

$\boldsymbol{v}_t$ に関する条件から、最適解において速度場はポテンシャルの勾配で与えられることがわかる。またこのとき、p_t に関する条件から次を得る：

$$\partial_t\phi_t + \frac{1}{2}|\nabla\phi_t|^2 = 0. \tag{46}$$

この方程式は解析力学における Hamilton–Jacobi 方程式

$$\partial_t S_t(\boldsymbol{x}) + H(\boldsymbol{x}, \nabla S_t(\boldsymbol{x})) = 0 \tag{47}$$

において、自由粒子のハミルトニアン $H(\boldsymbol{x},\boldsymbol{p}) = \frac{1}{2}|\boldsymbol{p}|^2$ を考えた場合に相当するため Hamilton–Jacobi 方程式と呼ばれる。

1.3.6 Benamou–Brenier 公式の双対性

1.2 章においては元々の最適輸送問題に対する双対問題を考えた。一方、同様にして式 (26) に対する双対問題を次のように与えられる：

$$\begin{aligned} &\inf_{\{p_t\},\{\boldsymbol{v}_t\}} \frac{1}{2}\int_0^1\int p_t(\boldsymbol{x})|\boldsymbol{v}_t(\boldsymbol{x})|^2 dxdt \quad \text{s.t.} \quad \partial_t p_t = -\nabla\cdot(p_t\boldsymbol{v}_t),\ p_0 = p^{(0)},\ p_1 = p^{(1)} \\ &= \sup_{\{\psi_t\}} \int (\psi_1(\boldsymbol{x})p^{(1)}(\boldsymbol{x}) - \psi_0(\boldsymbol{x})p^{(0)}(\boldsymbol{x}))dx \quad \text{s.t.} \quad \partial_t\psi_t + \frac{1}{2}|\nabla\psi_t|^2 \le 0. \end{aligned} \tag{48}$$

右辺の条件は微分不等式と呼ばれるが、先に見た通り最適解においては等号が達成される。この不等式は次のように現れる。まず Lagrange 未定乗数 2ψ を考えた時、汎函数の中に $2p_t(-\partial_t\psi_t + \frac{1}{2}|\boldsymbol{v}_t|^2 - \boldsymbol{v}_t\cdot\nabla\psi_t)$ という項が現れる。$p_t \ge 0$ を操作することでこれをいくらでも小さくできてしまわないためには、$-\partial_t\psi_t + \frac{1}{2}|\boldsymbol{v}_t|^2 - \boldsymbol{v}_t\cdot\nabla\psi_t \ge 0$ が成り立つ必要がある。右辺は二次関数で、$\boldsymbol{v}_t = \nabla\psi_t$ のとき最小化されるので、結局 $\partial_t\psi_t + \frac{1}{2}|\nabla\psi_t|^2 \le 0$ が成り立つ必要があることがわかる。この不等号は等号に置き換えられる。より一般のコスト関数に関しても、対応する Hamilton–Jacobi 方程式を用いて同様の関係が成り立つ。[3, Prop.5.48] ないし [2, Thm.7.36] 参照。

2　離散系における Wasserstein 距離 I

確率分布は連続変数だけのものではない。むしろ、実際の実験から得られるデータはヒストグラムにように離散的である。以下では変数の値の集合 X は離散的であるとし、簡単のため X は有限集合とする。有限集合 A に対し、$|A|$ で元の個数を表す。いま X は離散的な集合なので、規格化条件は $\sum_{x\in X} p_x^{(0)} = \sum_{x\in X} p_x^{(1)} = 1$ と表される。

2.1　グラフ理論

離散分布間の Wasserstein 距離の話を理解する上では、$x \in X$ を「離散的な状態」とみなすとわかりやすい。たとえば、スピン配位のようなもの $X = \{0,1\}^N$ が考えられる。このとき、一度にフリップできるスピンの個数の上限を定めれば、ある配位 $x \in X$ から遷移できる配位は限定される。離散分布間の Wasserstein 距離は、このような状態間の「遷移可能性」をベースに議論される。

このとき、グラフ理論的な考え方や、線形代数的グラフ理論の手法が有用である。状態 $x \in X$ をグラフの**ノード**（頂点）と同一視しよう。そして、ある状態 x から y への遷移が可能である場合、x と y との間に**有向エッジ** $(x,y) \in E_{\rightarrow}$ が存在すると定義する。以下では、あらゆる遷移は可逆であるとし、$e = (x,y) \in E_{\rightarrow}$ のみを以て $-e = (y,x)$ の存在をも表すことにする。よって $-e = (y,x) \notin E_{\rightarrow}$ とする[21]。組 $(X, E_{\rightarrow})$ を有向グラフと呼ぶ。

グラフは連結であるとする。すなわち、任意の $x, y \in X$ に対し、列 $P = (x_1, x_2, \dots, x_k)$ が存在して、$x_1 = x, x_k = y$ を満たし、全ての $i = 1, 2, \dots, k-1$ について (x_i, x_{i+1}) ないし (x_{i+1}, x_i) が $E_{\rightarrow}$ に属するものとする。始点終点にこだわらず、最後の条件のみを満たす列を一般に**経路**と呼び[22]、さらに最初の 2 つの条件を満たすものを x と y を結ぶ経路と呼ぶ。$k-1$ を P の経路長と呼ぶ。

有向グラフに対し、**接続行列** B を次で定義する。すなわち、$e = (x,y)$ のとき $B_{ie} = \delta_{iy} - \delta_{ix}$ とする。つまり、状態 i が e の終点ならば $B_{ie} = 1$, 始点ならば -1, エッジ e に含まれていなければ 0 ということである。このとき B は $|X| \times |E_{\rightarrow}|$ の行列となり、一般に非正方行列となる。また、グラフが連結ならば $\operatorname{rank} B = |X| - 1$ となることが知られている（証明は補遺 B）。

2.2　輸送コストとして

連続分布の場合と同様に、状態（位置）x から y へ砂山を運ぶコストを c_{xy} とし、輸送計画 π_{xy} の総コストを

$$\mathcal{C}[\pi] = \sum_{x,y} c_{xy}\pi_{xy} \tag{49}$$

と与えることができる。そして、状態 $p^{(0)}$, $p^{(1)}$ 間のカップリングの集合は

$$\Pi(p^{(0)}, p^{(1)}) = \left\{ \pi \mid \sum_y \pi_{xy} = p_x^{(0)},\ \sum_x \pi_{xy} = p_y^{(1)},\ \pi_{xy} \geq 0 \right\} \tag{50}$$

によって定められる。

具体的なコストを与えるため、グラフ上の距離関数 d_{xy} を用いる。d_{xy} を与えるためには、x と y をつなぐすべての経路を考える。今グラフは連結なので、経路長が最小のものが存在するから、その経路長をもって d_{xy} を定めることにする。たとえば図 3 の場合、$d_{12} = 3, d_{13} = 2, d_{23} = 1$ である。このとき d_{xy} は距離の公理を満たすことが示せる。

2.2.1　1-Wasserstein 距離

1-Wasserstein 距離が次で定義できる：

$$\mathcal{W}_1(p^{(0)}, p^{(1)}) = \min_{\pi \in \Pi(p^{(0)}, p^{(1)})} \sum_{x,y} d_{xy}\pi_{xy}. \tag{51}$$

[21] 物理的には、一つの状態遷移は複数の経路を持ちうる。例えば、あるスピン配位から別のスピン配位へ、熱浴とエネルギーの交換を伴って変化する時、熱浴の選択肢が複数あれば、対応する変化はグラフ上で区別される。以下では簡単のためそのような重複はないとして証明を行うが、$e = (x, y; 1), e' = (x, y; 2)$ のように、始点・終点に加えて追加のラベルの三つ組でエッジを指定することにより常に一般化できる。

[22] （承前）重複するエッジが存在する時、経路の定義に追加のラベルの情報も含める必要がある。

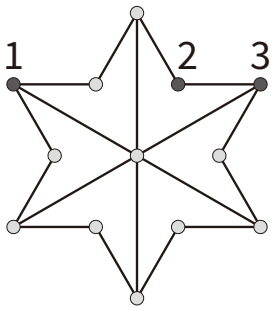

図 3　グラフの例。$d_{12}=3$, $d_{13}=2$, $d_{23}=1$ となる。

d_{xy} が距離関数であることから、この量もまた距離関数になることが直ちに従う。

2.2.2　2-Wasserstein 距離の問題

同様に、2-Wasserstein 距離が次で定義されそうである。

$$\mathcal{W}_2(p^{(0)},p^{(1)})=\min_{\pi\in\Pi(p^{(0)},p^{(1)})}\left[\sum_{x,y}d_{xy}^2\pi_{xy}\right]^{1/2}. \tag{52}$$

しかし、この距離は微分幾何学的な性質が悪く、微分が発散してしまう。このことを示すために、$X=\{a,b\}$, $E=\{(a,b)\}$ の場合を考え、分布を $p_a(\beta)=(1-\beta)/2$, $p_b(\beta)=(1+\beta)/2$ とパラメータづけする[23]（$\beta\in[-1,1]$）。$p^{(0)}=p(\beta_0)$, $p^{(1)}=p(\beta_1)$ とすると、式 (52) は次のように与えられる[24]：$\mathcal{W}_2(p(\beta_0),p(\beta_1))=\sqrt{|\beta_0-\beta_1|/2}$. さらに $\beta_0=\beta$, $\beta_1=\beta+d\beta$ とすれば、

$$\frac{\mathcal{W}_2(p(\beta),p(\beta+d\beta))}{d\beta}=\frac{1}{\sqrt{2}}\frac{\sqrt{|d\beta|}}{d\beta}\to\infty \tag{53}$$

となり、微分が発散してしまう。

2.3　Kantorovich 双対性

前章の双対性の節 1.2 は分布の滑らかさを必要としていなかったため、すべての結果がそのまま成り立つ。

まず、式 (13) と同様の Kantorovich 双対性が成り立つ：

$$\min_{\pi\in\Pi(p^{(0)},p^{(1)})}\sum_{x,y}c_{xy}\pi_{xy}=\max_{\phi,\psi}\sum_y\phi_y p_y^{(1)}-\sum_x\psi_x p_x^{(0)}\quad\text{s.t.}\quad\phi_y-\psi_x\le c_{xy}. \tag{54}$$

このとき相補性が存在し、最適解では $\pi_{xy}(c_{xy}-\phi_y+\psi_x)=0$ が常に成り立つ。

特に、$c_{xy}=d_{xy}$ とすると、式 (24) と同じく、次の Kantorovich–Rubinstein 双対性が成り立つ：

$$\min_{\pi\in\Pi(p^{(0)},p^{(1)})}\sum_{x,y}d_{xy}\pi_{xy}=\max_{\phi\|\phi\|_{\mathrm{Lip}}\le1}\sum_x\phi_x(p_x^{(1)}-p_x^{(0)}), \tag{55}$$

ここで

$$\|\phi\|_{\mathrm{Lip}}=\max_{x,y|x\neq y}\frac{|\phi_x-\phi_y|}{d_{xy}} \tag{56}$$

[23] 文献 [6] より。
[24] それなりに頑張る。

である。三角不等式から次のように表すこともできる[25]：

$$\|\phi\|_{\mathrm{Lip}} = \max_{(x,y)\in E_{\to}} |\phi_x - \phi_y|. \tag{57}$$

期待値を用いれば、

$$\mathcal{W}_1(p^{(0)}, p^{(1)}) = \max_{\phi \mid \|\phi\|_{\mathrm{Lip}} \le 1} [\langle \phi \rangle_1 - \langle \phi \rangle_0] = \max_{\phi} \frac{\langle \phi \rangle_1 - \langle \phi \rangle_0}{\|\phi\|_{\mathrm{Lip}}} \tag{58}$$

と書くことができる。

以下は離散系特有の結果である。L^1 距離は $L(p,q) = \sum_x |p_x - q_x|$ で定義されるのであった。このとき、式 (58) において $\phi_x = \mathrm{sgn}(p_x^{(1)} - p_x^{(0)})$ とすると、次の大小関係が明らかになる：

$$\mathcal{W}_1(p^{(0)}, p^{(1)}) \ge \frac{1}{2} L(p^{(0)}, p^{(1)}). \tag{59}$$

Lipschitz 条件は、接続行列を用いると次のように表される：$\max_{e\in E_{\to}} |[B^{\mathrm{T}}\phi]_e| \le 1$. 実際、

$$[B^{\mathrm{T}}\phi]_{(x,y)} = \sum_z (\delta_{yz} - \delta_{xz})\phi_z = \phi_y - \phi_x \tag{60}$$

となる。

2.4 運動論的な定式化

1.3 節では、分布の滑らかさを仮定したが、1-Wasserstein 距離の場合分布が滑らかでなくとも Benamou–Brenier 公式のようなものが与えられる。すなわち、次が成り立つ：

$$\mathcal{W}_1(p^{(0)}, p^{(1)}) = \min_J \sum_{e\in E_{\to}} |J_e| \quad \text{s.t.} \quad p^{(1)} - p^{(0)} = BJ. \tag{61}$$

最小化の条件の右辺を詳細に見ると、第 x 成分は次のような形で与えられる：

$$[BJ]_x = \sum_{(y,z)\in E_{\to}} (\delta_{xz} - \delta_{xy}) J_{yz} = \sum_{y\mid (y,x)\in E_{\to}} J_{yx} - \sum_{z\mid (x,z)\in E_{\to}} J_{xz}. \tag{62}$$

J_{xy} を x から y への流れと見做せば、最右辺の第一項は x への流入を、第二項は x からの流出を表す。したがって最小化の条件式は、「J はエッジ上の流れで、$p^{(0)}$ から $p^{(1)}$ への変化を表すものでなければならない」ということを意味することがわかる。

積分に関する三角不等式から、次のように表すこともできる：

$$\mathcal{W}_1(p^{(0)}, p^{(1)}) = \min_{\{p_t\},\{j_t\}} \int_0^1 \sum_{e\in E_{\to}} |j_{t,e}| dt \quad \text{s.t.} \quad \frac{dp_t}{dt} = Bj_t, \;\; p_0 = p^{(0)}, \;\; p_1 = p^{(1)}. \tag{63}$$

このことは以下のように簡単に理解できる。まず、式 (63) の最適化条件を満たす j_t に対し、$J = \int_0^1 j_t dt$ とすると、J は式 (61) の条件を満たす。このとき、積分の性質から

$$|J_e| = \left| \int_0^1 j_{t,e} dt \right| \le \int_0^1 |j_{t,e}| dt \tag{64}$$

が成り立つ。また逆に、任意の J に対して $j_{t,e} = J_e$ とすれば、式 (63) の目的関数は式 (61) の目的関数と同じ値を返す。よってこれら二つの最適化問題は同じ値を与えることがわかる。

この場合も時間幅に関する不定性があり、単位時間幅 $[0,1]$ を $[0,\tau]$ に読み替えても良い。この時 τ 倍の因子は現れず、単純に積分範囲だけが変更される。

---

[25] x, y が最短経路 $P = (x_1, \dots, x_k)$ によって結ばれる場合、$|\phi_x - \phi_y| = |\phi_{x_1} - \phi_{x_2} + \phi_{x_2} - \cdots + \phi_{x_{k-1}} - \phi_{x_k}| \le \sum_{i=1}^{k-1} |\phi_{x_i} - \phi_{x_{i+1}}|$ なので $|\phi_x - \phi_y|/d_{xy} \le \sum_{i=1}^{k-1} |\phi_{x_i} - \phi_{x_{i+1}}|/(k-1) \le \max_i |\phi_{x_i} - \phi_{x_{i+1}}|$ となる。任意の $e \in E_{\to}$ は何らかの最短経路の一部であるから $\max_{x,y} |\phi_x - \phi_y|/d_{xy} \le \max_{(x,y)\in E_{\to}} |\phi_x - \phi_y|$ が従う。一方、$(x,y) \in E_{\to}$ の時 $|\phi_x - \phi_y| = |\phi_x - \phi_y|/d_{xy}$ なので、$E_{\to}$ での最大化は $X \times X$ 内での最大化の一部分のみを考えていることになる。したがって $\max_{x,y} |\phi_x - \phi_y|/d_{xy} \ge \max_{(x,y)\in E_{\to}} |\phi_x - \phi_y|$ となる。よって等号が成り立つ。

2.4.1 接続行列

式 (61) の証明を行う前に、接続行列の一般的な性質について説明を行う。先に述べた通り、グラフが連結ならば $\operatorname{rank} B = |X| - 1$ となることが知られている。B は $|X| \times |E_{\to}|$ の行列だったから、左ゼロベクトル n が存在して $n^{\mathrm{T}}B = 0$ となることがわかる。この n は、全ての要素が 1 のベクトルの定数倍になることが知られている。言い方を変えると、任意の e に関して $\sum_x B_{xe} = 0$ が成り立つ（補遺 B 参照）。また、式 (60) および式 (62) は次のようにまとめられる：

$$[B^{\mathrm{T}}\phi]_{(x,y)} = \phi_y - \phi_x, \quad [-BJ]_x = \sum_{z|(x,z)\in E_{\to}} J_{xz} - \sum_{y|(y,x)\in E_{\to}} J_{yx}. \tag{65}$$

さて、これらの関係を通じて微分演算子 ∇ との類似性を考察しよう。式 (65) の第一式から $B^{\mathrm{T}}\phi$ がエッジ上での ϕ の差分を与え、勾配 $\nabla\phi$ と似た性質を持つことがわかる。また第二式から、$-BJ$ が状態からの総流出を与え、発散 $\nabla \cdot \boldsymbol{J}$ と対応づくことがわかる。部分積分の境界項を無視すると

$$\int \phi(\boldsymbol{x}) \nabla \cdot J(\boldsymbol{x}) dx = -\int \nabla\phi(\boldsymbol{x}) \cdot J(\boldsymbol{x}) dx \tag{66}$$

が成り立つが、$\int f(\boldsymbol{x})g(\boldsymbol{x})dx$ を関数の内積、$\int \boldsymbol{u}(\boldsymbol{x}) \cdot \boldsymbol{v}(\boldsymbol{x})dx$ をベクトル場の内積と考えるとここから $\mathrm{div}^{\dagger} = -\mathrm{grad}$ を得る。この関係は接続行列において $[-B^{\mathrm{T}}]^{\mathrm{T}} = -B$ と表現できる。そして、境界が無視できる場合、Gauss の定理から標語的に $\int dx \nabla \cdot = 0$ が成り立つ。この関係は $n^{\mathrm{T}}B = 0$ に対応している。

2.4.2 式 (61) の証明

以下の方針で証明を行う（永山龍那氏のアイデア）。まず、式 (61) の右辺が左辺の下限を与えることを示す。次に、式 (61) の右辺が Kantorovich–Rubinstein 双対性 (55) の右辺の上限を与えることを示す。そこで、以下のように定義を行う：

$$\begin{aligned} \mathcal{A}[J] &= \sum_{e\in E_{\to}} |J_e|, \quad \mathcal{A}^* = \min_{J|BJ=p^{(1)}-p^{(0)}} \mathcal{A}[J] \\ \mathcal{K}[\phi] &= \sum_x \phi_x (p_x^{(1)} - p_x^{(0)}), \quad \mathcal{K}^* = \max_{\phi|\|\phi\|_{\mathrm{Lip}}\le 1} \mathcal{K}[\phi]. \end{aligned} \tag{67}$$

すなわち、$\mathcal{W}_1(p^{(0)}, p^{(1)}) \ge \mathcal{A}^*$ および $\mathcal{A}^* \ge \mathcal{K}^*$ を示し、$\mathcal{W}_1(p^{(0)}, p^{(1)}) = \mathcal{K}^*$（式 (55)）から $\mathcal{W}_1(p^{(0)}, p^{(1)}) = \mathcal{A}^*$ を示す。

はじめに、$\delta(e, x, y)$ $(e \in E_{\to}, x, y \in X)$ を以下のように定義する：x から y をつなぐ最短経路を $P = (x_1, x_2, \dots, x_k)$ とする。ただし $x_1 = x$, $x_k = y$ であり、$d_{xy} = k - 1$ である。このとき、

$$\delta(e, x, y) = \begin{cases} 1 & \exists i,\ e = (x_i, x_{i+1}) \\ -1 & \exists i,\ e = (x_{i+1}, x_i) \\ 0 & \text{otherwise} \end{cases} \tag{68}$$

と定める。この関数は次の性質を満たす：$\sum_{e\in E_{\to}} |\delta(e, x, y)| = d_{xy}$. なぜなら、$P$ が持つエッジは d_{xy} 個であり、その全てが $E_{\to}$ に含まれているからである。また、次を満たす：$\sum_{e\in E_{\to}} B_{xe}\delta(e, y, z) = \delta_{yx} - \delta_{zx}$. これは以下のように示される。まず x が y と z をつなぐ最短経路 $P = (x_1, \dots, x_k)$ 上にない場合、各 $B_{xe}\delta(e, y, z)$ は常に 0 である。そして端点でない場合、ある $1 < i < k$ に関して $x = x_i$ となるが、$e_i = (x_{i-1}, x_i)$, $e_{i+1} = (x_i, x_{i+1})$ としたとき、$e \in E_{\to}$ に対して $B_{x(-e)} \coloneqq -B_{xe}$ とすれば

$$B_{xe_i}\delta(e_i, y, z) + B_{xe_{i+1}}\delta(e_{i+1}, y, z) = 1 - 1 = 0 \tag{69}$$

が成り立つ。その他の e について各 $B_{xe}\delta(e, y, z)$ がゼロになることは P の最短性から明らかである。最後に $x = y$ の場合 $e = (x, x_2)$ のみが、$x = z$ の場合 $e = (x_{k-1}, x)$ のみが非ゼロの寄与を与えることから符号が定まる。

さて、$\mathcal{W}_1(p^{(0)}, p^{(1)}) \ge \mathcal{A}^*$ を示そう。そのために、任意のカップリング π に対し $\mathcal{C}[\pi] \ge \mathcal{A}[J]$ および $BJ = p^{(1)} - p^{(0)}$ を満たす J を構成する。そのような J は次のように与えることができる：

$$J_e = \sum_{x,y} \delta(e, x, y)\pi_{xy}. \tag{70}$$

はじめに $BJ = p^{(1)} - p^{(0)}$ を示す。これは上に示した性質 $\sum_{e\in E_\rightarrow} B_{xe}\delta(e,y,z) = \delta_{yx} - \delta_{zx}$ から次のように直ちに従う：

$$[BJ]_x = \sum_{e\in E_\rightarrow} B_{xe} \sum_{y,z} \delta(e,y,z)\pi_{yz} = \sum_{y,z} \pi_{yz}(\delta_{yx} - \delta_{zx}) = \sum_z \pi_{xz} - \sum_y \pi_{yx} = p_x^{(1)} - p_x^{(0)}.$$

次に $\mathcal{C}[\pi] \geq \mathcal{A}[J]$ を示す。これは $\sum_{e\in E_\rightarrow} |\delta(e,x,y)| = d_{xy}$ から次のように示される：

$$\sum_{x,y} d_{xy}\pi_{xy} = \sum_{x,y}\sum_{e\in E_\rightarrow} |\delta(e,x,y)|\pi_{xy} = \sum_{e\in E_\rightarrow}\sum_{x,y} |\delta(e,x,y)|\pi_{xy} \geq \sum_{e\in E_\rightarrow} \left|\sum_{x,y} \delta(e,x,y)\pi_{xy}\right| = \sum_{e\in E_\rightarrow} |J_e|. \tag{71}$$

よって、$\mathcal{W}_1(p^{(0)}, p^{(1)}) \geq \mathcal{A}^*$ が示された。

次に、$\mathcal{A}^* \geq \mathcal{K}^*$ を示す。$BJ = p^{(1)} - p^{(0)}$ を満たす J および $\|\phi\|_{\mathrm{Lip}} \geq 1$ 満たす ϕ について、選び方によらずに不等式 $\mathcal{A}[J] \geq \mathcal{K}[\phi]$ が成り立つことを示す。そのために、1-Lipschitz 条件が $\forall e \in E_\rightarrow$, $|[B^{\mathrm{T}}\phi]_e| \leq 1$ と表現できたことを思い出す。すると

$$\sum_{e\in E_\rightarrow} |J_e| \geq \sum_{e\in E_\rightarrow} |[B^{\mathrm{T}}\phi]_e||J_e| = \sum_{e\in E_\rightarrow} |[B^{\mathrm{T}}\phi]_e J_e| \geq \left|\sum_x \sum_{e\in E_\rightarrow} \phi_x B_{xe} J_e\right| = \left|\sum_x \phi_x (p_x^{(1)} - p_x^{(0)})\right| \tag{72}$$

となり、所望の不等式が示される。よって $\mathcal{A}^* \geq \mathcal{K}^*$ が成り立つことがわかる。さらに以上の議論をまとめると、$\mathcal{W}_1(p^{(0)}, p^{(1)}) = \mathcal{A}^*$ を得る。

3　離散系における Wasserstein 距離 II

連続系の場合、2-Wasserstein 距離は解析的な性質が良く、以下でも概説するように勾配流が Fokker–Planck 方程式を導く等、物理的なダイナミクスとの親和性が高い。一方、前章で見た通り、離散系の場合 2-Wasserstein 距離は振る舞いが悪く、通常真剣な考察対象とは見做されない。その代わりに数学者が考案したのが、（究極的には）次のような形である：

$$\mathcal{W}_{\mathrm{WM}}(p^{(0)}, p^{(1)}) = \inf_{\{p_t\},\{f_t\}} \left[\int_0^1 \|f_t\|^2_{L(p_t)} dt\right]^{1/2} \quad \text{s.t.} \quad p_0 = p^{(0)},\ \ p_1 = p^{(1)},\ \ \frac{dp_t}{dt} = BL(p_t)f_t, \tag{73}$$

ただしここで、$L(p)$ は $|E_\rightarrow| \times |E_\rightarrow|$ の行列で、$e = (x,y)$ としたとき、

$$L_{ee'}(p) = \frac{w_e p_x - w_{-e} p_y}{\ln[w_e p_x/(w_{-e}p_y)]}\delta_{ee'} \tag{74}$$

のように正定数 $w_{\pm e}$ を用いて与えられる。一方 $f \in \mathbb{R}^{|E_\rightarrow|}$ であり、$\|f\|^2_{L(p)} = f^{\mathrm{T}}L(p)f$ である[26]。あとで示すように $L(p)$ は $\forall x,\ p_x > 0$ の場合正定値である。連続分布に対する Benamou–Brenier 公式 (26) および付随する連続の式 (27) と比べると、f が速度場に、$L(p)$ が確率分布に対応していることがわかる。この距離を便宜のため Wasserstein–Maas 距離縮めて WM 距離と呼ぼう[27]。WM 距離は勾配流方程式として次を与える：

$$\frac{dp_t}{dt} = BJ(p_t) \ \text{ with } \ J_e(p) = w_e p_x - w_{-e} p_y \ \ (e = (x,y)) \tag{75}$$

あるいは等価な表現

$$\frac{dp_{t,x}}{dt} = \sum_{y|(y,x)\,\mathrm{or}\,(x,y)\in E_\rightarrow} (w_{(y,x)}p_y - w_{(x,y)}p_x). \tag{76}$$

これらの方程式はマスター方程式と呼ばれ、離散変数の確率ダイナミクスの一般的な表現である。以下、はじめにモチベーションとなる勾配流方程式の概要について説明したのち、WM 距離の性質を解説する。

[26] w_e を状態間遷移の遷移レートと捉えると、$L(p)$ は確率流 $J_e(p)$ と熱力学力 $F_e(p) = \ln[w_e p_x/(w_{-e}p_y)]$ とを結びつける行列であると解釈できる。

[27] J. Maas が文献 [6] において導入したため。一般的な呼称ではない。

3.1 Wasserstein 勾配流方程式

3.1.1 連続系の場合

実空間における勾配 $\nabla f(\boldsymbol{x})$ は $\boldsymbol{x}_0 = \boldsymbol{x}$ を満たす任意の時間発展 $\boldsymbol{x}_t$ を用いて $\frac{d}{dt} f(\boldsymbol{x}_t)|_{t=0} = \dot{\boldsymbol{x}}_0 \cdot \nabla f(\boldsymbol{x})$ を計算することによって特徴づけられる。確率分布の空間における「勾配」もまた、$\dot{p}_0$ を変化させることで計算できることが期待される。ところで実空間の場合、内積は Euclid 内積によって与えられたが、確率分布の微分を相手にする際にどうすべきかは自明ではない。

そこで通常用いられるのが次のような接空間と内積の構造である。なお以下では連続分布を念頭におく。まず、確率分布の時間発展 p_t に対して $\dot{p}_t = -\nabla \cdot (p_t \nabla \phi)$ を満たすポテンシャル ϕ が一意的に[28]存在することに注目する。そして、M を確率分布の多様体とし、$p \in M$ における接空間 $T_p M$ を

$$T_p M = \{-\nabla \cdot (p \nabla \phi) \mid \phi \in (\text{適当な空間})\} \tag{77}$$

によって定義する[29]。たとえば $p_0 = p$ を満たす時間発展 $\{p_t\}$ は $\dot{p}_0 \in T_p M$ となる。そして、$u = -\nabla \cdot (p \nabla \phi)$, $v = -\nabla \cdot (p \nabla \psi)$ に対する内積を

$$\langle u, v \rangle_p = \int p(\boldsymbol{x}) \nabla \phi(\boldsymbol{x}) \cdot \nabla \psi(\boldsymbol{x}) d\boldsymbol{x} \tag{78}$$

によって定義する。

このとき、確率分布の関数 $F(p)$ の勾配は次のように定義できる：$\operatorname{grad} F(p)$ は $p_0 = p$ を満たす任意の時間発展 p_t に対して、

$$\frac{d}{dt} F(p_t)|_{t=0} = \langle \dot{p}_0, \operatorname{grad} F(p) \rangle_p \tag{79}$$

を満たし、$-\nabla \cdot (p \nabla \phi_F)$ の形で表される関数として与えられる

この内積は Wasserstein 距離と整合的である。式 (78) の形は、すでに Benamou–Brenier 公式のポテンシャルによる表現 (33) に現れていた。すなわち、$\|\dot{p}\|_p = \sqrt{\langle \dot{p}, \dot{p} \rangle_p}$ とすれば、式 (33) は

$$\mathcal{W}_2(p^{(0)}, p^{(1)})^2 = \inf_{\{p_t\}} \int_0^1 \|\dot{p}_t\|_{p_t}^2 dt \tag{80}$$

と表すことができる。

勾配流方程式を考えるため、$F(p) = D(p \| p^{\mathrm{eq}})$ の場合を考える。ただしここで

$$D(p \| q) = \int p(\boldsymbol{x}) \ln \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} d\boldsymbol{x} \tag{81}$$

は Kullback–Leibler ダイバージェンスであり、

$$p^{\mathrm{eq}}(\boldsymbol{x}) = \frac{1}{Z} e^{-\beta U(\boldsymbol{x})} \tag{82}$$

は平衡分布で、β は逆温度、$U(\boldsymbol{x})$ はポテンシャル関数に対応し、$Z = \int e^{-\beta U(\boldsymbol{x})} d\boldsymbol{x}$ は規格化因子である。このとき、ある p_t に対し、$p_0 = p$, $\dot{p}_0 = -\nabla \cdot (p \nabla \phi)$ とすると、

$$\frac{d}{dt} D(p_t \| p^{\mathrm{eq}})|_{t=0} = \int \dot{p}_0 (\ln p - \ln p^{\mathrm{eq}}) d\boldsymbol{x} \tag{83}$$

$$= \int -\nabla \cdot (p \nabla \phi)(\ln p + \beta U + \ln Z) d\boldsymbol{x} \tag{84}$$

$$= \int p \nabla \phi \cdot \nabla (\ln p + \beta U) d\boldsymbol{x} \tag{85}$$

[28] 適切な境界条件の元、定数分の足し引きを無視して。

[29] より正確な主張については [2, Chapter 8.1.2] など参照。

を得る。ただし一行目で規格化 $\int \dot{p} dx = 0$ を用い、三行目で部分積分を行う際、表面項を無視した。よって

$$\operatorname{grad} D(p\|p^{\mathrm{eq}}) = -\nabla \cdot [p\nabla(\beta U + \ln p)] = \nabla \cdot [(-\beta \nabla U)p] - \nabla^2 p \tag{86}$$

を得る。対応する勾配流方程式

$$\partial_t p_t = -\operatorname{grad} D(p_t\|p^{\mathrm{eq}}) \tag{87}$$

は、逆温度 β、拡散係数 1 の白色ガウスノイズの下、ポテンシャル力 $-\nabla U(\boldsymbol{x})$ を受けるブラウン粒子の運動を表す Fokker–Planck 方程式を与える（補遺 A も参照）。

3.1.2 離散系の場合

Fokker–Planck 方程式は連続変数の確率的時間発展を与える。離散変数における対応物として、状態 x から y への遷移が、無限小時間 dt において確率 $w_{xy}dt$ で生じるダイナミクスを考えると、確率分布の時間発展は次のマスター方程式によって与えられる：

$$\frac{dp_{t,x}}{dt} = \sum_y (w_{yx}p_{t,y} - w_{xy}p_{t,x}). \tag{88}$$

ここでは可逆性 $w_{xy} > 0 \iff w_{yx} > 0$ を仮定し、$w_{xy} > 0$ のときエッジ (x, y) を設定することにする。ただし $(x, y) \in E_{\to}$ のとき $(y, x) \notin E_{\to}$ とする。すると、式 (88) は式 (76) ないし式 (75) に一致する。

接続行列 B と ∇ の関係に鑑みつつ、天下り的に離散確率分布上の接空間と内積を以下のように定める。すなわち、

$$T_pM = \{BL(p)B^{\mathrm{T}}\phi \mid \phi \in \mathbb{R}^{|X|}\} \tag{89}$$

とする。$\dot{p} = BL(p)B^{\mathrm{T}}\phi$ を満たす ϕ は常に存在し[30]、$\dot{p} \in T_pM$ となることがわかる。$u = BL(p)B^{\mathrm{T}}\phi$, $v = BL(p)B^{\mathrm{T}}\psi$ に対して内積を

$$\langle u, v\rangle_p = \phi^{\mathrm{T}}BL(p)B^{\mathrm{T}}\psi \tag{90}$$

で定める。あとで示す WM 距離のポテンシャルを用いた表式 (96) と比べると、この内積は WM 距離と整合的であることがわかり、対応するノルムを $\|\cdot\|_p$ と書くと

$$\mathcal{W}_{\mathrm{WM}}(p^{(0)}, p^{(1)}) = \inf_{\{p_t\}}\left[\int_0^1 \|\dot{p}_t\|_{p_t}^2 dt\right]^{1/2} \quad \text{s.t.} \quad p_0 = p^{(0)},\ \ p_1 = p^{(1)} \tag{91}$$

となる。ただし式 (80) が 2-Wasserstein 距離と T_pM の内積とを結びつける一方、この式は本質的には右辺が左辺を定義しているに過ぎない点には注意が必要である。

$D(p\|p^{\mathrm{eq}})$ の勾配を求めよう。ただしここで $D(p\|q) = \sum_x p_x \ln(p_x/q_x)$ であり、p^{eq} は平衡分布で、$(x, y) \in E_{\to}$ に対し $w_{xy}p_x^{\mathrm{eq}} = w_{yx}p_y^{\mathrm{eq}}$ を満たす（そのような p^{eq} の存在を w に対して課す）。このとき $\dot{p}_0 = BL(p)B^{\mathrm{T}}\phi$ に対し、

$$\frac{d}{dt}D(p_t\|p^{\mathrm{eq}})|_{t=0} = \sum_x \dot{p}_{t,x} \ln \frac{p_{t,x}}{p_x^{\mathrm{eq}}} = \phi^{\mathrm{T}}BL(p)B^{\mathrm{T}}\mu \quad (\mu_x = \ln(p_x/p_x^{\mathrm{eq}})) \tag{92}$$

となる。したがって、$\operatorname{grad} D(p\|p^{\mathrm{eq}}) = BL(p)B^{\mathrm{T}}\mu$ となる。ここで、平衡分布が存在する際成り立つ以下の関係に注意する：$e = (x, y)$ のとき

$$L_{ee}(p) = \frac{w_{xy}p_x - w_{yx}p_y}{\ln[w_{xy}p_x/(w_{yx}p_y)]} = \frac{w_{xy}p_x - w_{yx}p_y}{\ln[p_xp_y^{\mathrm{eq}}/(p_yp_x^{\mathrm{eq}})]} = \frac{J_e(p)}{-[B^{\mathrm{T}}\mu]_e}. \tag{93}$$

したがって、$\operatorname{grad} D(p\|p^{\mathrm{eq}}) = -BJ(p)$ となり、$D(p\|p^{\mathrm{eq}})$ の勾配流方程式 $\dot{p}_t = -\operatorname{grad} D(p_t\|p^{\mathrm{eq}})$ がマスター方程式を与えることがわかる。

[30] まず $\sum_x \dot{p}_x = \frac{d}{dt}\sum_x p_x = 0$ であり、$\operatorname{rank} B = |X| - 1$ より $\dot{p} \in \operatorname{im} B$ となる。ただし im は行列の像。また、$L(p)$ はフルランクなので $\operatorname{rank} BL(p)B^{\mathrm{T}} = \operatorname{rank} B$. よって $\operatorname{im} B = \operatorname{im} BL(p)B^{\mathrm{T}}$ であり、$\dot{p} = BL(p)B^{\mathrm{T}}\phi$ を満たす ϕ が存在する。さらに、定数の足し引きを除いて一意であることもわかる。

以上の勾配流方程式の結果にとどまらず、MW 距離は連続系におけるいくつかの微分幾何学的関係性を離散分布に対しても再現する。たとえば、以下のように確率分布の曲がり具合の下限を定義することができる。初期および終分布を一組選び、式 (91) において右辺を最小化する経路 $\{p_t^*\}$ を取ってくる。このとき

$$D(p_t^*\|p^{\mathrm{eq}}) \le (1-t)D(p_0^*\|p^{\mathrm{eq}}) + tD(p_1^*\|p^{\mathrm{eq}}) - \frac{\kappa}{2}t(1-t)\mathcal{W}_{\mathrm{WM}}(p_0^*, p_1^*) \tag{94}$$

をどんな初期および終分布に対しても満たす κ が存在した時、そのうち最大のものによって「Ricci 曲率」が定義できる。この定義は、連続系においてさまざまな関数不等式や、Fokker–Planck 方程式の固有値との関係が知られており、興味深いと言える。もっとも、距離が「遷移レート」w を含み、なんらかの「ダイナミクス」の情報を必要とすることは注意を要する。後述するように、「最適輸送」というコンセプトとの物理的関係付けに関しても必ずしも簡単ではない。

3.2 MW 距離の性質

3.2.1 距離関数

MW 距離はその名の通り距離関数となる。式 (5) の条件のうち、1–3 については対数平均の性質から容易に示すことができるが、三角不等式は証明がいささか厄介である。興味のある読者は文献 [7] の V.A. 節を参照されたい。

3.2.2 時間幅に関する不定性

連続の場合と同様に、時間幅に関する不定性が存在する。すなわち次が成り立つ：

$$\mathcal{W}_{\mathrm{WM}}(p^{(0)}, p^{(1)}) = \inf_{\{p_t\},\{f_t\}} \left[\tau \int_0^\tau \|f_t\|^2_{L(p_t)} dt\right]^{1/2} \quad \text{s.t.} \quad p_0 = p^{(0)},\ \ p_\tau = p^{(1)},\ \ \frac{dp_t}{dt} = BL(p_t)f_t. \tag{95}$$

このことを確かめるには、$\dot{p}_t = BL(p_t)f_t$ のとき、$t' = \tau t$, $p'_{t'} = p_{t'/\tau}$, $f'_{t'} = f_{t'/\tau}/\tau$ が $\dot{p}'_{t'} = BL(p'_{t'})f'_{t'}$ を満たすことを考えれば良い。

3.2.3 ポテンシャルによる表現

先にも触れた通り、式 (73) はポテンシャルを用いて次のように表すことができる：

$$\mathcal{W}_{\mathrm{WM}}(p^{(0)}, p^{(1)}) = \inf_{\{p_t\},\{\phi_t\}} \left[\int_0^1 \|B^{\mathrm{T}}\phi_t\|^2_{L(p_t)} dt\right]^{1/2} \quad \text{s.t.} \quad p_0 = p^{(0)},\ \ p_1 = p^{(1)},\ \ \frac{dp_t}{dt} = BL(p_t)B^{\mathrm{T}}\phi_t. \tag{96}$$

B^{T} と勾配の対応を思い出すと、この式は連続の場合の式 (33) の離散対応になっており、同様にして示すことができる。

すなわち、$\dot{p}_t = BL(p_t)f_t$ を満たす f_t に対し、$\dot{p}_t = BL(p_t)B^{\mathrm{T}}\phi_t$ を満たす ϕ_t を用いて $f_t = B^{\mathrm{T}}\phi_t + g_t$ としたとき、成分 g_t は（コストの観点から）無駄であることを示せる。実際、

$$\|f_t\|^2_{L(p_t)} = \|B^{\mathrm{T}}\phi_t\|^2_{L(p_t)} + \|g_t\|^2_{L(p_t)} + 2(B^{\mathrm{T}}\phi_t)^{\mathrm{T}}L(p_t)g_t \tag{97}$$

$$= \|B^{\mathrm{T}}\phi_t\|^2_{L(p_t)} + \|g_t\|^2_{L(p_t)} + 2\phi_t^{\mathrm{T}}BL(p_t)g_t \tag{98}$$

であるが、$BL(p_t)g_t = BL(p_t)f_t - BL(p_t)B^{\mathrm{T}}\phi_t = 0$ なので、

$$\|f_t\|^2_{L(p_t)} = \|B^{\mathrm{T}}\phi_t\|^2_{L(p_t)} + \|g_t\|^2_{L(p_t)} \ge \|B^{\mathrm{T}}\phi_t\|^2_{L(p_t)} \tag{99}$$

を得る。ここで $J_t^{\mathrm{c}} := L(p_t)g_t$ が満たす関係 $BJ_t^{\mathrm{c}} = 0$ は、補遺 B において説明されている通り、グラフ上のサイクルを表す。

3.2.4 双対性

双対性についてはやや変更が加わり、次が成り立つ：

$$\begin{aligned}&\inf_{\{p_t\},\{f_t\}} \frac{1}{2}\int_0^1 \|f_t(\boldsymbol{x})\|^2_{L(p_t)} dt \quad \text{s.t.} \quad p_0 = p^{(0)},\ \ p_\tau = p^{(1)},\ \ \frac{dp_t}{dt} = BL(p_t)f_t \\ &= \sup_{\{\psi_t\}} (\psi_1^{\mathrm{T}}p^{(1)} - \psi_0^{\mathrm{T}}p^{(0)}) \quad \text{s.t.} \quad q^{\mathrm{T}}\frac{d}{dt}\psi_t + \frac{1}{2}\|B^{\mathrm{T}}\psi_t\|^2_{L(q)} \le 0,\ \forall q \in \Sigma_X,\end{aligned} \tag{100}$$

ただし $\Sigma_X = \{q \in \mathbb{R}^{|X|} \mid \sum_x q_x = 1,\ q_x \geq 0\}$ である。ここで不等式が奇妙な形をしているが、$B^{\mathrm{T}}\psi \leftrightarrow \nabla\psi$ および $L(p) \leftrightarrow p$ の対応関係を思い出すと奇妙ではなくなる。この関係を Kantorovich 双対性と呼ぶことがある。こちらも発見的な証明が文献 [7] の Appendix G にある。

3.3 最適輸送？

WM 距離と物理的な輸送との関係はあまり議論されない。連続系の場合、Monge 問題の最適輸送写像と Benamou–Brenier 公式の速度場が実は同一のものを見ていたのに対して、離散系ではそもそも Monge 問題のような定式化が難しい。なぜなら、輸送を座標間の関係だけで決めてしまうと、各点で重みが厳密に保存されてしまうためである。たとえば、分布 $(1/8, 3/4, 1/8)$ を $(1/4, 1/2, 1/4)$ に移すようなことができない。一方連続系では、平均 0 分散 1 の一変数 Gauss 分布を、分散 4 の Gauss 分布に移す場合、$T(x) = 2x$ によって引き伸ばすことで輸送が可能である。このことは非可算無限集合の柔軟性によるが、離散系ではそのような柔軟性が存在しない。

一方、カップリングによる定式化はすでに見た通り $r = 2$ とすると微分的な性質が悪くなるため、WM 距離と本質的に異なる概念であることが示唆される。したがって、素朴な意味では物理的な輸送と相性はよろしくない。

しかしながら、連続の式の形が出てくる以上、なんらかのダイナミクスとの関係は示唆されるはずである。実際、f ないし ϕ がなにか直接的にコントロールできる量で、連続の式を通じて確率の変化を与える場合、式 (73) および (96) は $p^{(0)}$ から $p^{(1)}$ への「最適」な動かし方を与えることができる。ただしここで「最適」とかっこ付けしたのは、このときコストの解釈が自明ではないからである。式 (73) および (96) において積分されている量をなぜ小さくしたいのかについては議論の余地が残る。

一方、f をゆらぎの熱力学における熱力学力と解釈した場合、コストの解釈はエントロピー生成として自明になる。式 (95) の表式に拠れば、有限時間幅 τ におけるエントロピー生成に時間幅 τ をかけたものの平方根をコストとすることになる。エントロピー生成とはエネルギーの無駄遣いのことであり、これは物理的に自然なコストと言える。しかしながら、このとき以下に見るように連続の式に問題が生じる。

熱力学力とは遷移レート w（w_{xy} はいま $x \to y$ の遷移レート）および確率分布 p を用いて

$$F_{(x,y)} = \ln \frac{w_{xy} p_x}{w_{yx} p_y} \tag{101}$$

と定義される。すなわち、L 同様遷移レートに直接依存することになる。このことを明示した上で連続の式（マスター方程式）を書くと次のようになる：

$$\frac{dp_t}{dt} = BL(p_t, w)F(p_t, w). \tag{102}$$

w を動かせば F は好きなだけ動かせるから、式 (73) は次のようにも書き表すことができる

$$\mathcal{W}_{\mathrm{WM}}(p^{(0)}, p^{(1)}) = \inf_{\{p_t\},\{w_t\}} \left[\int_0^1 \|F(p_t, w_t)\|^2_{L(p_t,w)} dt \right]^{1/2} \quad \text{s.t.}\ \ p_0 = p^{(0)},\ \ p_1 = p^{(1)},\ \ \frac{dp_t}{dt} = BL(p_t, w)F(p_t, w_t). \tag{103}$$

この表式だと、L を決める w が固定されているため、もはや連続の式は物理的意味を失ってしまっているようにも思えるが、これは問題ではない。

いま、$F_{(x,y)}(p, w)$ は w_{xy} と w_{yx} の両方ではなく、それらの比にしか依存していない。そのため、$L(p, w)$ を保ったまま w を動かすだけで、$F(p, w)$ を自在に変化させることが可能である。このことは、w が $2|E_{\to}|$ 個のパラメータである一方、$L(p, w)$ は $|E_{\to}|$ 次元の対角行列であり、$F(p, w)$ も $|E_{\to}|$ 次元のベクトルであることを考慮すれば明らかになる。すなわち、次が成り立つ：ある w に対し、

$$\begin{aligned} \mathcal{W}_{\mathrm{WM}}(p^{(0)}, p^{(1)}) &= \inf_{\{p_t\},\{w_t\}} \left[\int_0^1 \|F(p_t, w_t)\|^2_{L(p_t,w_t)} dt \right]^{1/2} \\ \text{s.t.}\ \ p_0 &= p^{(0)},\ \ p_1 = p^{(1)},\ \ \frac{dp_t}{dt} = BL(p_t, w_t)F(p_t, w_t),\ \ L(p_t, w_t) = L(p_t, w). \end{aligned} \tag{104}$$

つまり、遷移レート w_t は、各時刻において $L(p_t, w_t) = L(p_t, w)$ が成り立つ限り好きにいじって良い。しかしこの条件は、通常実現するのが難しいのではなかろうか。

補遺 A　Fokker–Planck 方程式との関係性

「はじめに」で述べたように、2-Wasserstein 距離は Brown 運動のようなダイナミクスと相性が良い。「Brown 運動のようなダイナミクス」を一般に Langevin ダイナミクスと呼ぶが、ここでは特に最適輸送理論との関係が明確な overdamped Langevin ダイナミクスおよびその熱力学的性質について概説する。詳しくは文献 [8, 9] や、フリーだが有用なノート [10, 11] などを参照。

変数 $\boldsymbol{x}$ が overdamped Langevin ダイナミクスに従う時、その時間発展 $\boldsymbol{x}_t$ は次の確率微分方程式を満たす：

$$d\boldsymbol{x}_t = D\beta \boldsymbol{f}_t(\boldsymbol{x}_t)dt + \sqrt{2D}d\boldsymbol{W}_t. \tag{105}$$

両辺を dt で割れば、通常の微分方程式のようにも読める。$\boldsymbol{f}_t$ は力学的な力で、たとえば外的なポテンシャル $U(\boldsymbol{x})$ が働く場合、$\boldsymbol{f}_t = -\nabla U$ となる。D は拡散係数、β は熱浴の逆温度である。$d\boldsymbol{W}_t$ は熱浴によるランダムな力を表し、$\langle dW_{i,t}\rangle = 0, \langle dW_{i,t}dW_{j,t}\rangle = \delta_{ij}dt$ を満たし、$|t-t'| > dt$ の場合 $d\boldsymbol{W}_t$ と $d\boldsymbol{W}_{t'}$ とが独立になる[31]。$d\boldsymbol{W}_t$ は $\sqrt{dt}$ のオーダーの量なので、dt で両辺を割る際は注意が必要である（$\boldsymbol{\xi}_t = \frac{d\boldsymbol{W}_t}{dt}$ とすると、$\boldsymbol{\xi}_t$ はデルタ関数的な分散を持つことになる）。

そのほかにもいくつか注意する点がある。運動方程式は、本来時間に関する 2 階微分方程式であるが、その 2 階微分の項を無視することによってこの 1 階の微分方程式は得られる。ここで温度が出てくるのは、熱浴によるランダムな力の強さ D と、元々の 2 階の運動方程式に現れる摩擦力 $-\gamma\dot{\boldsymbol{x}}$ の強さとが、熱浴が平衡状態であるために揺動散逸関係 $\gamma^{-1} = D\beta$ によって結びつくからである。

さてこのとき、適当な初期分布 p_0 から出発した変数が、時刻 t において値 $\boldsymbol{x}$ を取る確率 $p_t(\boldsymbol{x})$ は、次の Fokker–Planck 方程式に従うことが知られている：

$$\partial_t p_t(\boldsymbol{x}) = -\nabla\cdot(p_t(\boldsymbol{x})\boldsymbol{v}_t(\boldsymbol{x})), \quad \boldsymbol{v}_t(\boldsymbol{x}) = D\beta\boldsymbol{f}_t(\boldsymbol{x}) - D\nabla\ln p_t(\boldsymbol{x}), \tag{106}$$

確率微分方程式と $\boldsymbol{v}_t(\boldsymbol{x})$ との右辺らを見比べると、ランダムな力 $\sqrt{2D}d\boldsymbol{W}_t$ が Fokker–Planck 方程式では「決定論的な」項 $-D\nabla\ln p_t(\boldsymbol{x})$ に化けていることがわかる。この項は結局 Fokker–Planck 方程式において $D\nabla^2 p_t(\boldsymbol{x})$ となるので、拡散を表現していることがわかる。

この系において熱力学を考えるには、熱浴が受け取る「熱」を定義する必要がある。そのためにしばしば採用される仮説が、次の定義である：

$$dQ_t = \boldsymbol{f}_t(\boldsymbol{x}_t)\circ d\boldsymbol{x}_t. \tag{107}$$

ここで $\circ$ は Stratonovich 積と呼ばれる積で、$\boldsymbol{f}_t((\boldsymbol{x}_t + \boldsymbol{x}_{t+dt})/2)d\boldsymbol{x}_t$ によって定義される。このような定義が意味を持つのは、$d\boldsymbol{x}_t$ に含まれる $d\boldsymbol{W}_t$ が、2 乗しても $o(dt)$ のオーダーにならず、無視できないことが一因となる[32]。

加えて、「系」すなわち自由度 $\boldsymbol{x}$ のエントロピーを、Shannon エントロピーに Boltzmann 定数 k_B をかけた $-k_\mathrm{B}\ln p_t(\boldsymbol{x}_t)$ によって定義すれば、確率的な全系のエントロピー変化 $d\Sigma_t$ は次のように与えられる：

$$d\Sigma_t = d(-k_\mathrm{B}\ln p_t(\boldsymbol{x}_t)) + k_\mathrm{B}\beta dQ_t. \tag{108}$$

この平均を計算すると、結局 $\langle d\Sigma_t\rangle$ は dt のオーダーになることがわかり、エントロピー生成レート $\dot{\Sigma}_t = \langle d\Sigma_t\rangle/dt$ は次で与えられることが示される：

$$\dot{\Sigma}_t = \frac{k_\mathrm{B}}{D}\int p_t(\boldsymbol{x})|\boldsymbol{v}_t(\boldsymbol{x})|^2 dx. \tag{109}$$

有限時間幅 $[0,\tau]$ におけるエントロピー生成はこれを積分することによって

$$\Sigma_\tau = \frac{k_\mathrm{B}}{D}\int_0^\tau\int p_t(\boldsymbol{x})|\boldsymbol{v}_t(\boldsymbol{x})|^2 dxdt \tag{110}$$

[31] 実際には、確率微分方程式は初めに時間を細切れにするので、時間幅が被るような状況には普通ならないと思われる。
[32] Riemann 積分の定義において、代表点の任意性があったこととは対照的である。

と与えられる。

2-Wasserstein 距離の Benamou–Brenier 公式によれば

$$\mathcal{W}_2(p_0,p_\tau)^2 = \inf_{\{q_t\},\{\boldsymbol{u}_t\}} \tau \int_0^\tau \int q_t(\boldsymbol{x})|\boldsymbol{u}_t(\boldsymbol{x})|^2 dx dt \ \text{ s.t. } \partial_t q_t(\boldsymbol{x}) = -\nabla\cdot(q_t(\boldsymbol{x})\boldsymbol{u}_t(\boldsymbol{x})),\ q_0 = p_0,\ q_\tau = p_\tau \tag{111}$$

が成り立つのだった。ただしここで時間幅に関する不定性から有限時間幅 τ を設け、実際のダイナミクスと区別するために $q, \boldsymbol{u}$ という記号を用いた。したがって、次の不等式が成り立つ：

$$\Sigma_\tau \geq \frac{k_\mathrm{B}}{D\tau}\mathcal{W}_2(p_0,p_\tau)^2. \tag{112}$$

この不等式は τ への下限として

$$\tau \geq \frac{k_\mathrm{B}\mathcal{W}_2(p_0,p_\tau)^2}{D\Sigma_\tau} \tag{113}$$

あるいは

$$\tau \geq \sqrt{\frac{k_\mathrm{B}}{D\langle\dot{\Sigma}\rangle_\tau}}\mathcal{W}_2(p_0,p_\tau) \tag{114}$$

と書き改められる。ここで $\langle g\rangle_\tau = \frac{1}{\tau}\int_0^\tau g_t dt$ である。すなわち、距離 $\mathcal{W}_2(p_0,p_\tau)$ だけ離れた状態間の遷移にかかる時間 τ には、その間のエントロピー生成ないし平均エントロピー生成レートに反比例する下限が存在することが示される。

この等号を達成するプロトコルは、最適輸送理論の知識を使うと明示的に与えられる。Benamou–Brenier 公式 (111) における最適な $\boldsymbol{u}$ は、あるポテンシャル ϕ によって $\boldsymbol{u}_t = \nabla\phi_t$ と与えられるのだった（1.3.2 節参照）。そして、この ϕ_t は Hamilton–Jacobi 方程式を満たす（1.3.5 節参照）。これに加えて p_0 と p_τ を結ぶ連続の式を解けば、最適な時間発展が求まる [12]。このとき、Fokker–Planck 方程式と Benamou–Brenier 公式の間の類似性に着目すれば、

$$D\beta\boldsymbol{f}_t(\boldsymbol{x}) - D\nabla\ln p_t(\boldsymbol{x}) = \nabla\phi_t(\boldsymbol{x}) \tag{115}$$

とすればよいことがわかるので、

$$U_t(\boldsymbol{x}) = -\frac{\phi_t(\boldsymbol{x}) + D\ln p_t(\boldsymbol{x})}{D\beta} \tag{116}$$

というポテンシャルを加えれば最適な輸送が行えることがわかる。

補遺 B　接続行列のランク・サイクルとの関係

以下では、連結なグラフ $(X, E_\to)$ の接続行列 B のランクが $|X|-1$ になることを示す[33]。なお連結でない場合、行列は連結な部分グラフの接続行列の直和になる。

まず初めに具体的なグラフを通じて接続行列 B に親しもう。図 4 では、$X = \{1,2,3,4\}$ の 4 状態が存在し、$E_\to = \{(1,2),(1,3),(1,4),(2,3),(3,4)\}$ の 5 つの遷移が存在している。したがって接続行列は 4×5 の行列になる。

第一行は $(-1,-1,-1,0,0)$ であり、エッジ $e=(1,2),(1,3),(1,4)$ がノード $x=1$ から出ており、$e=(2,3),(3,4)$ は関わりがないことを表す。第一列は $(-1,1,0,0)^\mathrm{T}$ であり、エッジ $e=(1,2)$ がノード $x=1$ から出て、$x=2$ に入っていることに対応している。

各列がエッジに対応することを思えば、各列は始点と終点のノード（行）において -1 と 1 をそれぞれ一つずつ含むことになる。その他の成分は 0 なのだから、$\sum_x B_{xe} = 1+(-1) = 0$ が成り立つ。したがって $\operatorname{rank} B \leq |X|-1$ である。

[33] 文献 [13] にもっと簡単な証明があることを教えてもらったが、この議論だとサイクルや全域木の感じにも触れられるので、やや面倒な証明を行なっている。なお筆者はこの本にめちゃくちゃお世話になった。

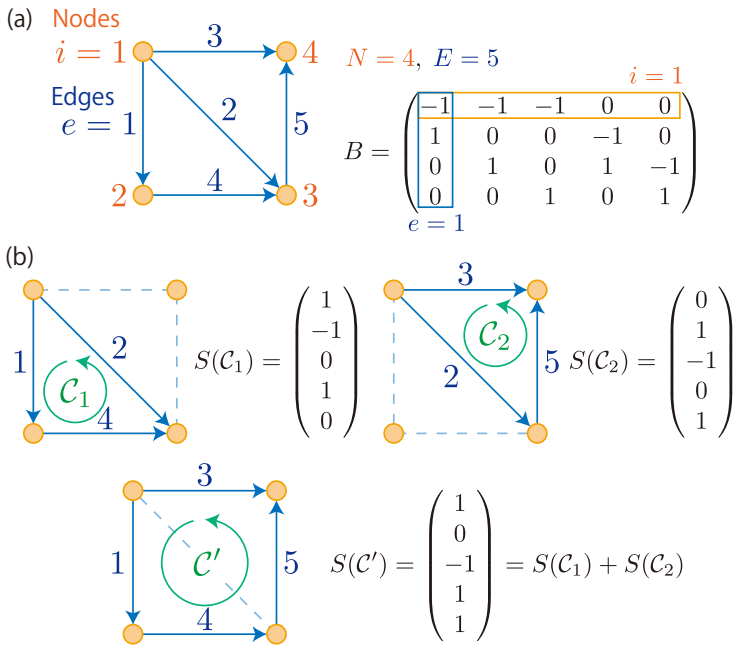

図 4　文献 [7] より転載。i は x と読み替えられたい。$N=|X|, E=|E_{\to}|$ である。

この等号が成り立つことを確認するために、全域木という概念を導入しよう。全域木とは簡単に言ってしまえば元のグラフの一部（部分グラフ）で、全ノードを最小限のエッジでつなぐようなものである。したがって、$T=(X,E_T)$ が全域木である場合、$E_T \subset E_{\to}$ であり、T は連結である。例えば $E_T=\{(1,2),(1,3),(1,4)\}$ や $\{(1,2),(1,3),(3,4)\}$ や $\{(1,3),(1,4),(2,3)\}$ は全域木になる。このように全域木は一意ではないが、そのエッジの本数は一意に定まる。元のグラフが連結な場合、全域木のエッジの本数は $|X|-1$ になる。この結果はノードの個数に関する数学的帰納法によって示すことができる。直感的には、枝分かれがなければ一本鎖の植木算であり、枝分かれがある場合、剪定された枝ではノードの数＝エッジの数になることをイメージすれば正しそうに感じられる。

重要な性質として全域木はサイクルを含まない。なぜなら、サイクルからエッジを 1 本除いても連結性は損なわれないからである。そもそもサイクルとはなんだろうか。図 4 では、$1\to2\to3\to1$ というサイクルが存在する。エッジの言葉で言えば、$(1,2),(2,3)$ に関して $+1$ の、$(1,3)$ に対して -1 の重みをつければこのサイクルが表現できる。この重みを B の $1\tilde{3}$ 列に掛けて足せば、和は 0 になる。それもそのはず、B の各列は出入りの関係性を表していたのに対し、サイクルでは出入りの関係が一回りしてプラマイゼロになるのだから。この重みをまとめて $S(C_1)=(1,1,-1,0,0)^{\mathrm{T}}$ とベクトルにすれば、当然 $BS(C_1)=0$ が成り立つ。したがって、接続行列を用いると、サイクルは列同士の線形従属関係（あるいは B の右ゼロベクトル）として機械的に理解可能になる。

さて、全域木はサイクルを含まないのだった。すなわち、全域木において選ばれるエッジ E_T に注目すれば、それらを組み合わせて出入りの関係の帳尻を合わせることはどう頑張ってもできない。したがって、元のグラフの接続行列 B には、$|E_T|=|X|-1$ 個の独立な列が存在する。よって、$\operatorname{rank} B=|X|-1$ が成り立つ。

参考文献

[1] 佐藤竜馬. 最適輸送の理論とアルゴリズム. 講談社, 2023.

[2] Cédric Villani. *Optimal transport: old and new*, Vol. 338. Springer Berlin, Heidelberg, 2009.

[3] Cédric Villani. *Topics in optimal transportation*, Vol. 58. American Mathematical Soc., 2021.

[4] 寒野義博, 土谷隆. 最適化と変分法. 丸善, 2014.

[5] Muka Nakazato and Sosuke Ito. Geometrical aspects of entropy production in stochastic thermodynamics based on Wasserstein distance. *Phys. Rev. Res.*, Vol. 3, No. 4, p. 043093, 2021.

[6] Jan Maas. Gradient flows of the entropy for finite Markov chains. *J. Funct. Anal.*, Vol. 261, No. 8, pp.

2250–2292, 2011.
[7] Kohei Yoshimura, Artemy Kolchinsky, Andreas Dechant, and Sosuke Ito. Housekeeping and excess entropy production for general nonlinear dynamics. *Phys. Rev. Res.*, Vol. 5, No. 1, p. 013017, 2023.
[8] Ken Sekimoto. *Stochastic energetics*, Vol. 799. Springer, 2010.
[9] 齊藤圭司. ゆらぐ系の熱力学 非平衡統計力学の発展. SGC ライブラリ, No. 182. サイエンス社, 2022.
[10] 伊藤創祐. 非平衡科学（講義ノート）. [pdf].
[11] 金澤輝代士. 揺らぐ系の熱力学の基礎. [pdf].
[12] Erik Aurell, Carlos Mejía-Monasterio, and Paolo Muratore-Ginanneschi. Optimal protocols and optimal transport in stochastic thermodynamics. *Phys. Rev. Lett.*, Vol. 106, No. 25, p. 250601, 2011.
[13] 高崎金久. 線形代数とネットワーク. 日本評論社, 2017.