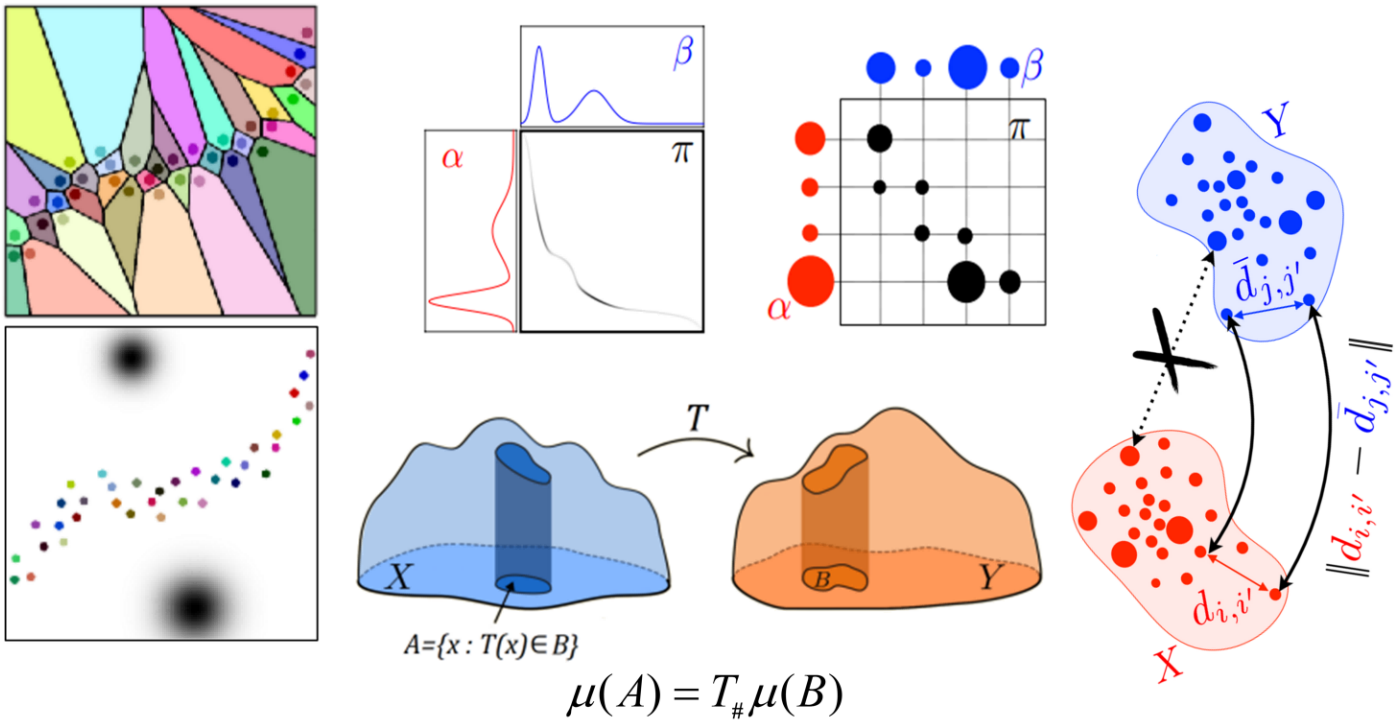
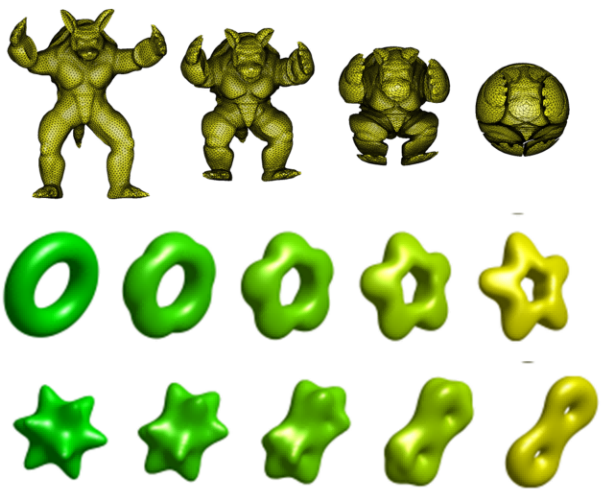
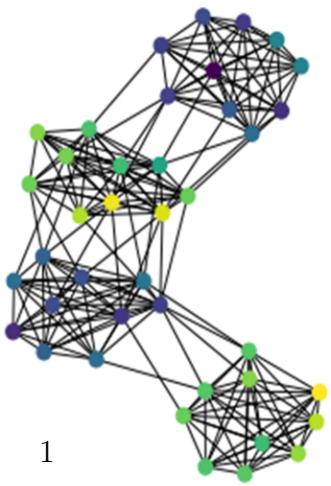
Note of Optimal Transport

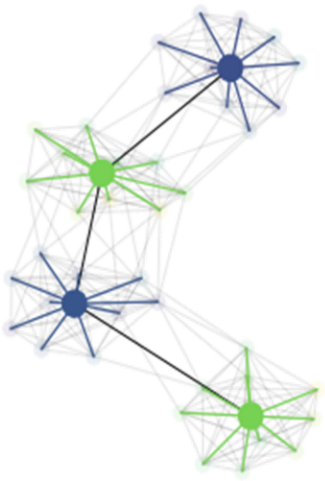
赵显文

2025 年 5 月 23 日









目录

1 引言

最优传输（1781），起源于法国大革命时期，提出者 Monge 曾在拿破仑手下任职，正是这个历史悠久的话题，在近代历久弥新。特别是近期，一个一度被深度学习范式主导学术界的学期，也算异军突起。最优传输，作为一个数学问题，沟通了流体力学、Monge-Ampere 方程等 PDE 理论与黎曼几何、Ricci 流等现代几何领域; 在经济学、运筹学、物理学（甚至有人用它来模拟宇宙进化过程）、生命科学、以及各类工程学，如生成式人工智能、域适应等人工智能领域、平均场博弈、向量场与流、图论、粒子运动或多体动力学等有重要应用。一大批知名学者研究它的各种方面，杰出人物如菲尔兹奖得主 Villani（2010）与 Figalli（2018），诺贝尔经济学的 Kantorovich（1975）等。值得一提的是，于 2006 年分别独立提出平均场博弈的研究者之一–菲尔兹奖得主 Lions(1994)–正是 Villani 的老师，而平均场博弈亦与最优传输关联密切。同样，非线性偏微分领域大家，因在自由边界问题和与最优传输关联的 Monge-Ampere 方程正则性理论有开创贡献而获阿贝尔奖的 Luis A. Caffarelli（卡法雷利，2012 年沃尔夫奖，2023 年 Abel 奖，当然其杰出工作还包含对 Navier-Stokes 方程正则性的研究），亦是 Figalli 的博后导师。此外，在最优传输领域“冠名”过的 Gromov 是几何大家（1993 年沃尔夫奖，2009 年阿贝尔奖）。

系统介绍最优传输的书籍包含 Villani 于 2003 的“Topics in Optimal Transportation”，和 2010 年的“Optimal Transport Old and New”（近千页），Santambrogio 于 2015 年的“Optimal Transport for Applied Mathematicians”，法国Peyré等人的 “Computational Optimal Transport” ，国内顾险峰（师从使用过蒙日安培方程解决卡拉比-丘猜想的菲尔兹奖得主丘成桐）于 2020 年的“最优传输理论与计算”,2021 年起出现了开源库POT，2023 年 Figalli 的“An Invitation to Optimal Transport, Wasserstein Distances, and Gradient Flows”，2024 年Sinho Chewi （Yale）、Jonathan Niles-Weed（NYU）与 Philippe Rigollet（MIT）的Statistical Optimal Transport。可以看出拥有约 250 年的最优传输理论在最近正蓬勃兴起，不论理论与应用。

参考上述书籍以及一些论文，本篇笔记仅就最优传输应用向一些广为人知的形式、算法进行整理或者摘抄，为个人使用！

2 经典最优传输

2.1 Monge and Kantorovich 问题

想象一下：现在有一堆沙子，我们要想用这堆沙子做一个艺术造型，比如一座山。那么我们需要按照山的形状将沙子运到对应的位置上，并且运输的过程沙子的量是保持的，也就是不会凭空产生和消失。如果是远距离原始，且沙子的量很大，那运输沙子是要花销的，我们总是想要以最经济的方式运送这些沙子。类似的，战争中的派兵布阵也是这样，战士由一些集中营运送到设计好阵型的阵地上，这个过程需要运输费用，我们的目标是想最小化运输成本。这些问题都可以建模为最优传输问题。最先规范化这个数学模型的是法国数学家 Monge，把战士从一个位置 x 运送到另一个位置 y, 用初始位置 x 标记每个战士，这形成了一个传输映射 $y = T(x)$，并假设所需的成本为 $c(x, y)$。这个运输过程中人不能凭空多出来、也不能突然消失，且每个人是一个整体不能被分割。这个问题的最小化运输成本，在形式化上表达为 Monge 问题：

$$(Monge\ Problem): \quad \min_{T} \int c(x, T(x))d\mu$$
$$s.t. \quad T_\sharp \mu = \nu$$

其中，$T : X \to Y$ 为传输映射（说明一个 x 对应一个 y, 运输过程 x 上的质量不能分割），μ, ν 分别为空间 X, Y 上的测度，可以简单视 $d\mu = f(x)dx$，其中 $f(x)$ 表示概率密度，$\mu(A)$ 可理解为集合 A 发生的概率,$\mu(A) = \int_A f(x)dx$。把 $T\sharp\ \mathcal{P}(X) \to \mathcal{P}(Y)$ 记为 *push forward*, 满足 $\forall B \subset Y$,$T\sharp\mu(B) = \mu(T^{-1}(B))$，有质量守恒之意.

Remark: 条件 $\forall B \subset Y$,$\mu(T^{-1}(B)) = \nu(B)$ 等价于对 $\forall$ 有界的 Borel 可测函数 h, $\int_Y h(y)d\nu(y) = \int_X h(T(x))d\mu(x)$. 这个等价关系的证明就利用实分析里示性函数、简单函数逼近连续函数那一套。有界性保证积分有意义（当然，本质只要 h^+ 或 h^- 的积分有一个有限即可）。

求解蒙日问题是很棘手的。一个重要问题是可行解 T 的空间 $\sum(\mu, \nu) = \{T \mid T_\sharp\mu = \nu\}$ 在弱收敛意义下是非闭的（非紧），这就导致求解过程中得到的（使传输代价下降的）传输映射序列不一定收敛到一个可行解（也不一定是一个映射）。当然了，也不一定是凸的，因而也不一定能求出全局最优；当然有最优解也不一定唯一

Kantorovich（诺贝尔经济学奖，提出线性规划）在苏联战时共产主义时期将 Monge 问题进行 relax（1942）。引入传输方案 $\gamma(x, y)$，将质量不可分的传输映射，变成质量可分的传输方案，可以简单理解为联合概率密度，且边缘分布分别为源域、目标域上的概率密度。Kantorovich 问题的规范化表述如下。

给定两个测度 $\mu \in \mathcal{P}(X), \nu \in \mathcal{P}(Y)$, 以及传输代价 $c : X \times Y \to [0, +\infty]$,Kantorovich 问题 (简记为 KP) 为求解:

$$(KP) \quad \inf_{\gamma \in \Pi(\mu,\nu)} \int_{X \times Y} c(x, y) d\gamma(x, y)$$

其中 $\Pi(\mu, \nu) := \{\gamma \in \mathcal{P}(X \times Y) \mid (\pi_x)_\sharp \gamma = \mu, (\pi_y)_\sharp \gamma = \nu\}$, π_x, π_y 分别表示向 X, Y 空间的投影映射。

对最优传输的计算研究，特别是计算机领域，也多集中在 Kantorovich 问题上。

作为应用,多采用离散形式。给定两个分布 $\boldsymbol{p} \in \sum_{n_1}, \boldsymbol{q} \in \sum_{n_2}$, 其中 $\sum_n = \{\boldsymbol{p} \in \mathbb{R}_+^n \mid \sum_{i=1}^n p_i = 1\}$ 为含有 n 个分区的直方图。$\Pi(\boldsymbol{p}, \boldsymbol{q}) = \{\boldsymbol{T} \in \mathbb{R}_+^{|\boldsymbol{p}| \times |\boldsymbol{q}|} \mid \boldsymbol{T}\mathbb{1}_{|\boldsymbol{q}|} = \boldsymbol{p}, \boldsymbol{T}^\intercal \mathbb{1}_{|\boldsymbol{p}|} = \boldsymbol{q}\}$, 则 Kantorovich 问题可转化为:

$$(KP) \quad \inf_{T \in \Pi(\boldsymbol{p},\boldsymbol{q})} < C, T >_F = \inf_{T \in \Pi(\boldsymbol{p},\boldsymbol{q})} \sum_{i,j} C_{i,j} T_{i,j} = \inf_{T \in \Pi(\boldsymbol{p},\boldsymbol{q})} \mathrm{Tr}(C^\top T) = \inf_{T \in \Pi(\boldsymbol{p},\boldsymbol{q})} \mathrm{Tr}(T^\top C)$$

展开来写，对应的 Wassertein 距离或者 Kantorvorvich 问题为:

$$\begin{aligned} \mathcal{W}(\mathbf{P}_1, \mathbf{P}_2) = \inf_{T \in \mathbb{R}^{n_1 \times n_2}} & \sum_{i.j} T_{i,j} \cdot C_{i,j} \\ \text{s.t.} \quad & T \geq 0 \\ & \sum_j T_{i,j} = \mu_i, \forall i \\ & \sum_i T_{i,j} = \nu_j, \forall j \end{aligned}$$

2.2 Duality

对偶理论形式的核心就是 "$(Primal) = \inf\sup \geq \sup\inf = (Dual)$"，当问题为凸优化问题（目标函数为凸函数，可行解集为凸集）时，对偶间隙为 0, 即前面的不等号变为等号。本小节仅介绍经典 Kantorovich 问题的对偶形式，其他形式如流体力学视角下的对偶形式将在介绍算法中自然引入。

先看一般的优化问题:

$$\begin{aligned} & \min_x f(x) \\ s.t. \quad & \begin{cases} g(x) \geq 0 \\ h(x) = 0 \end{cases} \end{aligned}$$

定义上述 Lagrange 函数 $\mathcal{L}(x, \lambda, \mu) = f(x) - u^\intercal g(x) - v^\intercal h(x)$,$u \succeq 0$. 可以发现:

$$\max_{u \succeq 0, v} \mathcal{L}(x, \lambda, \mu) = \begin{cases} f(x), & g(x) \geq 0, h(x) = 0 \\ +\infty, & others \end{cases}$$

因而，原问题等价于 $\min_x \max_{u\succeq 0,v} \mathcal{L}(x,\lambda,\mu)$. 交换 min 与 max 即为对偶问题：

$$(DP) \quad \max_{u\succeq 0,v} \min_x \mathcal{L}(x,\lambda,\mu) = \max_{u\succeq 0,v} \min_x [f(x) - u^{\mathsf{T}} g(x) - v^{\mathsf{T}} h(x)]$$

现在，回到 Kantorovich 问题上来。假设 $\gamma \geq 0$, 则 Kantorovich 问题的 Lagrange 函数为

- 连续形式

$$\mathcal{L}(\gamma,\phi,\psi) = \int_{X\times Y} C d\gamma - \int_X \phi d((\pi_x)_\sharp \gamma - \mu) - \int_Y \psi d((\pi_y)_\sharp \gamma - \nu)$$

- 离散形式

$$\mathcal{L}(T,\boldsymbol{f},\boldsymbol{g}) =< C, T > - < \boldsymbol{f}, T\mathbb{1}_{n_2} - \boldsymbol{\mu} > - < \boldsymbol{g}, T^{\mathsf{T}}\mathbb{1}_{n_1} - \boldsymbol{\nu} >$$

取 $\phi \in C_b(X), \psi \in C_b(Y)$，其中 $C_b(X)$ 表示定义在 X 上的有界连续函数空间, 则连续形式的 Kantorovich 问题的对偶为：

$$\begin{aligned}
(DP): & \sup_{\phi(x),\psi(y)} \inf_{\gamma} \left[\int_{X\times Y} C d\gamma - \int_X \phi d((\pi_x)_\sharp \gamma - \mu) - \int_Y \psi d((\pi_y)_\sharp \gamma - \nu)\right] \\
= & \sup_{\phi(x),\psi(y)} \inf_{\gamma} \left[\int_{X\times Y} C d\gamma + \int_X \phi(x) d\mu + \int_Y \psi(y) d\nu - \int_X \phi(x) d(\pi_x)_\sharp \gamma - \int_Y \psi(y) d(\pi_y)_\sharp \gamma\right] \\
= & \sup_{\phi(x),\psi(y)} \inf_{\gamma} \left[\int_{X\times Y} C d\gamma + \int_X \phi(x) d\mu + \int_Y \psi(y) d\nu - \int_{X\times Y} \phi(x) + \psi(y) d\gamma\right] \\
= & \sup_{\phi(x),\psi(y)} \left[\int_X \phi(x) d\mu + \int_Y \psi(y) d\nu + \inf_{\gamma} \int_{X\times Y} c(x,y) - \phi(x) - \psi(y) d\gamma\right]
\end{aligned}$$

由于要求 $\gamma \geq 0$, 有：

$$\inf_{\gamma\geq 0} \int_{X\times Y} c(x,y) - \phi(x) - \psi(y) d\gamma = \begin{cases} 0, & \phi \oplus \psi \leq c \\ -\infty, & \phi \oplus \psi > c \end{cases}$$

其中 $\oplus$ 表示逐点相加, 即 $\phi \oplus \psi(x,y) := \phi(x) + \psi(y)$。因此，对偶问题可进一步化简为：

$$(DP): \sup_{\substack{\phi(x),\psi(y) \\ \phi\oplus\psi\leq c}} \left[\int_X \phi(x) d\mu + \int_Y \psi(y) d\nu\right]$$

离散形式也可类似推导，当然也可以根据连续形式直接写出来：

$$(DP): \sup_{\substack{\boldsymbol{f}\in\mathbb{R}^{n_1}, \boldsymbol{g}\in\mathbb{R}^{n_2} \\ \boldsymbol{f}_i+\boldsymbol{g}_j\leq C_{i,j}}} [< \boldsymbol{f}, \boldsymbol{\mu} > + < \boldsymbol{g}, \boldsymbol{\nu} >] = \sup_{\substack{\boldsymbol{f}\in\mathbb{R}^{n_1}, \boldsymbol{g}\in\mathbb{R}^{n_2} \\ \boldsymbol{f}_i+\boldsymbol{g}_j\leq C_{i,j}}} \left[\sum_{i=1}^{n_1} \boldsymbol{f}_i \boldsymbol{\mu}_i + \sum_{j=1}^{n_2} \boldsymbol{g}_j \boldsymbol{\nu}_j\right]$$

Note1: 如果说原问题是最小化传输成本，那么对偶问题相当于雇佣公司去搬运。公司会有特定的搬运定价，$\phi(x)$ 表示从 x 点处搬运的费用，$\psi(y)$ 表示在 y 处卸载的费用。

公司想要最大化自己的收益。当然公司的定价时不会超过个人亲自搬运的费用，也就是 $\phi(x)+\psi(y)\leq c(x,y)$。

Note2: 从对偶问题的形式上来看，就是对 $\boldsymbol{f},\boldsymbol{g}$ 有一种 balance, 也就是 $\boldsymbol{f}+C,\boldsymbol{g}-C$ 都是可行的（C 这里可以是函数），也不改变目标函数的值，因此完全可以对 $\boldsymbol{f}$ 加个约束，比如直接 Lipschitz-1 的约束。

上述形式下的对偶问题，其可行域为 $\{\phi\in C_b(X),\psi\in C_b(Y)\mid \phi\oplus\psi\leq c\}$ 非紧，这就导致对 DK 问题解的存在性研究仍有一些困难。那就得引入新的结构 *c-transform*。

定义（c-transform）： 给定一个 X 上的函数 $\chi(x):X\to\bar{\mathbb{R}}$, 定义它的 c-transform 如下

$$\chi^c(y)=\inf_{x\in X}c(x,y)-\chi(x)$$

定义（$\bar{c}$-transform）： 给定一个 Y 上的函数 $\xi(x):Y\to\bar{\mathbb{R}}$, 定义它的 $\bar{c}$-transform 如下

$$\xi^{\bar{c}}(x)=\inf_{y\in Y}c(x,y)-\xi(y)$$

Note: 上述两个变换定义本质上是相同的，只不过根据定义域做了区分。

定义（c-concave）： 称一个在 X 上的函数 ϕ 是 c-concave, 如果存在一个 Y 上函数 ξ,s.t.,$\phi(x)=\xi^{\bar{c}}(x)$,c-concave 函数的集合记为 c-conc(X)

定义（$\bar{c}$-concave）： 称一个在 Y 上的函数 ψ 是 $\bar{c}$-concave, 如果存在一个 X 上函数 χ,s.t.,$\psi(y)=\chi^c(y)$,$\bar{c}$-concave 函数的集合记为 $\bar{c}$-conc(Y)

关于 Kantorovich 原问题、对偶问题解的存在性（数学描述）以及两者的等价（从凸优化来看确实等价，不过可以对最优传输问题具体化）请参考引言中提到的多本参考书，或者剑桥大学的一篇最优传输 Note。

性质 1（c-transform）： $\phi^{c\bar{c}}\geq\phi$, 等号成立当且仅当 ϕ 为 c-concave，通常情况下 $\phi^{c\bar{c}}$ 是大于 ϕ 的最小 c-concave 函数。同理把上述 c 换成 $\bar{c}$,$\bar{c}$ 换成 c 亦成立

证明： 不等号：由定义有 $\phi^{c\bar{c}}(x)=\inf_y c(x,y)-\phi^c(y)=\inf_y\{c(x,y)-\inf_{\tilde{x}}[c(\tilde{x},y)-\phi(\tilde{x})]\}$. 由于 $\inf_{\tilde{x}}[c(\tilde{x},y)-\phi(\tilde{x})]\leq c(x,y)-\phi(x)$，因此

$$\phi^{c\bar{c}}\geqslant\inf_y\{c(x,y)-c(x,y)+\phi(x)\}=\phi(x)$$

. 同理可得 $\psi^{\bar{c}c}(y)\geq\psi(y)$.

若 $\phi(x)$ 为 c-concave 函数，则 $\exists\xi\ Y\to\bar{\mathbb{R}},s.t.,\phi(x)=\xi^{\bar{c}}(x)$,

$$\phi=\xi^{\bar{c}}\Rightarrow\phi^c=\xi^{\bar{c}c}\geqslant\xi\Rightarrow\varphi^{c\bar{c}}\leqslant\xi^{\bar{c}}=\phi$$

最后一个"⇒" 是因为 $\bar{c}-$ 变换的定义（要减去对应函数）.

最小性: 任取 c-concave 函数 $\tilde{\phi}(x) = \chi^{\bar{c}}(x)$, 且 $\tilde{\phi}(x) \geq \phi(x), \forall x \in X$. 则有: $\chi^{\bar{c}} \geqslant \phi \Rightarrow \chi^{\bar{c}c} \leqslant \phi^c \Rightarrow \chi \leqslant \phi^c \Rightarrow \tilde{\phi} = \chi^{\bar{c}} \geqslant \phi^{c\bar{c}}$ ■

由此将产生一个推论:

推论 (c-transform): $\forall \phi(x) : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}(: \bar{\mathbb{R}})$。同理 $\forall \psi(y) : \mathcal{Y} \to \bar{\mathbb{R}}$ 则 $\phi^{c\bar{c}c} = \phi^c$, 则 $\psi^{\bar{c}c\bar{c}} = \psi^{\bar{c}}$

推论的证明即利用 ϕ^c 为 $\bar{c}$-concave 函数, $\psi^{\bar{c}}$ 为 c-concave 由性质立得。

定义 (legendre 对偶): 函数 $f : X \to Y$ 的 Legendre 对偶为 $f^*(y) = \sup_{x \in X} < x, y > -f(x)$。即一堆线性函数的上包络.

上述其实是广义的 Legendre 对偶, 最开始的定义在可微函数上的 Legendre 对偶为为 $f^*(p) = xp - f(x), p = f'(x)$. 本质上 legender 对偶为 $f(x)$ 在 x 处切线的负截距。x=r 处的切线方程: $y = f(r) + (x - r)f'(r) = xf'(r) - (rf'(r) - f(r))$.

性质 2 (Legendre 对偶): 函数 f 的 legendre 对偶 f^* 为凸函数。反过来也成立:

命题 (凸函数): 函数 $f : \mathbb{R}^n \to \mathbb{R}$ 为凸函数且下半连续, 当且仅当 $\exists g$ 满足 $f = g^*$, 也当且仅当 $f^{**} = f$. 对于一般函数 f 而言,f^{**} 为函数的凸化, $f \geq f^{**}$, 等号成立当且仅当 f 为凸函数. 一般而言,f^{**} 为小于函数 f 的最大凸函数.

性质 3 (Legendre 对偶): 互为 Legendre 对偶的函数, 其导函数互为逆映射。

Note: c-transform 可以看作 Legendre 的一种泛化。f^{**} 可以看作函数 f 的凸化, 是不大于原函数的最大凸函数。类似的 $\phi^{c\bar{c}}$ 是大于原函数 ϕ 最小的 c-concave 函数。在下一节的 Brenier 定理我们会讨论 $c(x, y) = \frac{1}{2}\|x - y\|^2$ 下两者的关系, 从最差传输问题来看, 此时 c-transform 就是 Legendre 变换

现在, 我们借助 c-transform 继续研究 Kantorovich 的 dual problem.

命题: 对于上述对偶问题的任意可行解 $(\phi(x), \psi(y))$, 则 $(\phi(x), \phi^c(y))$ 亦可行且更优, 同理 $(\psi^{\bar{c}}(x), \psi(y))$ 亦可行且更优。

由 c-concave 和 $\bar{c}$-concave 本质是一样的, 只是在最优传输背景下区分了函数的定义域, 因此 $\phi(x)$ 和 $\psi(y)$ 是对称的, 下面只证明前一个, 后一个把 x->y, y->x 即可。

证明: 可行性: 由于 $\phi^c(y) = \inf_{x \in X} c(x, y) - \phi(x)$, 因此 $\phi^c(y) \leq c(x, y) - \phi(x), \forall x \in X, y \in Y$, 即 $\phi^c(y) + \phi(x) \leq c(x, y), \forall x \in X, y \in Y$. 因此 $(\phi(x), \phi^c(y))$ 为 (DP) 问题的可行解

更优性: 由于 $(\phi(x), \psi(y))$ 为 (DP) 的可行解, 因此 $\psi(y) \leq c(x, y) - \phi(x), \forall x \in X, y \in Y$. 右侧对 x 取下确界有 $\psi(y) \leq \inf_{x \in X} c(x, y) - \phi(x) = \phi^c(y)$, 于是 $\int_Y \psi(y) d\nu \leq \int_Y \phi^c(y) d\nu$ ■.

推论: 任意可行解 $(\phi(x), \psi(y))$,$(\psi^{\bar{c}}(x), \psi^{\bar{c}c}(y))$ 亦可行且更优。同理 $(\phi^{c\bar{c}}(x), \phi^c(y))$

相比于 $(\phi(x), \psi(y))$ 亦可行且更优

证明： 基于上面的的命题，先获得更优可行解 $\psi^{\bar{c}}(x), \psi(y)$, 再次获得更优可行解 $(\psi^{\bar{c}}(x), \psi^{\bar{c}c}(y))$ 得证。当然还可以再用，不过结果都一样了:$(\psi^{\bar{c}c\bar{c}}(x), \psi^{\bar{c}c}(y))$ 根据上面的推论等于 $(\psi^{\bar{c}}(x), \psi^{\bar{c}c}(y))$ ■

在一点点符号误用下，对于上述的 $(\psi^{\bar{c}}(x), \psi^{\bar{c}c}(y))$, 如果记 $\phi(x) = \psi^{\bar{c}}$, 则可写成 $(\phi(x), \phi^c(x))$ 的形式。由于其可行且更优，那么对于原始 Kantorovich 问题的对偶问题:

$$(DP): \sup_{\substack{\phi(x),\psi(y) \\ \phi\oplus\psi\le c}} \left[\int_X \phi(x)d\mu + \int_Y \psi(y)d\nu\right]$$

可等价于如下形式:

$$(Semi\ DP-1): \sup_{\phi\in c-conc(X)} \left[\int_X \phi(x)d\mu + \int_Y \phi^c(y)d\nu\right]$$

或

$$(Semi\ DP-2): \sup_{\psi\in \bar{c}-conc(Y)} \left[\int_X \psi^{\bar{c}}(x)d\mu + \int_Y \psi(y)d\nu\right]$$

Remark: 若 (ϕ, ψ) 为（DP）问题的最优解，则 $\phi = \psi^{\bar{c}}, \psi = \phi^c$.

2.3 Brenier 理论与 Monge-Ampere Equation

定义（支撑集 (support))： 函数 $f\ X \to \bar{\mathbb{R}}$ 的支撑集为使得函数值大于 0 的点集的闭包,$i.e., supp(f) := \overline{\{x|f(x) > 0\}}$.

定义（c-subdifferential)： 函数 φ 的 c-次微分为

$$\partial^c\varphi = \{(x, y) \mid \varphi(z) \le \varphi(x) + (c(z, y) - c(x, y)), \forall z \in X\}$$

c-次微分的定义完全类比函数的次梯度:

$$\partial f(\mathbf{x}) \equiv \{\mathbf{g} \in \mathbb{E}^* : f(\mathbf{y}) \ge f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x}\rangle \ for\ all\ \mathbf{y} \in X\}$$

命题（次微分的等价定义)： $\partial^c\varphi = \{(x, y) \mid \varphi(x) + \varphi^c(x) = c(x, y)\}$.

证明： 由次微分定义 $\varphi(z) \le \varphi(x) + (c(z, y) - c(x, y)), \forall z \in X$, 有 $c(x, y) - \varphi(x) \le c(z, y) - \varphi(z)\ \forall z \in X$, 右侧取下确界，有 $c(x, y) - \varphi(x) \le \varphi^c(y)$. 另一方面，由 $c-$tranform 的定义有 $\varphi^c(y) \le c(x, y) - \varphi(x)$, 因此 $c(x, y) - \varphi(x) \ge \varphi^c(y)$. ■

2.3.1 Brenier 定理

本节主要研究成本函数 $c(x,y)=h(x-y)$, h 为严格凸函数下的最优传输问题, 特别是 $h=\frac{1}{2}|x|^2$. 在这种情况下, 最优传输方案等价于最优传输映射, 也就是说 Kantorovich 问题的最优传输方案 γ 的形式为 $(id,T)_\sharp\mu$. 此时最优传输方案 γ 的支撑集为一条曲线。

由于 $\min(KP)=\max(DP)$, 有:

$$\int_{X\times Y}c(x,y)d\gamma=\int_X\phi d\mu+\int_Y\phi^c d\nu=\int_{X\times Y}\phi(x)\oplus\phi^c(y)d\gamma$$

可得: $\phi(x)\oplus\phi^c(y)=c(x,y)$ 在 $X\times Y$ 上 $a.e.$, 成立。设 $(x,y)\in supp(\gamma)$, 则 $\phi(x)+\phi^c(y)=c(x,y)$. 换言之最优传输方案的支撑集就属于最优 Kantorovich 势能函数 ϕ 的 *c-subdifferential*,i.e.,$supp(\gamma)\subseteq\partial^c\phi$.

有下述结论:

命题 ([1] 的 Proposition 1.15): 给定 C^1 光滑的代价函数 c,γ 是从 μ 到 ν 的最优传输方案 ,$\boldsymbol{\phi}$ 是相应的 Kantorovich 势能函数, 假设 (x_0,y_0) 属于 γ 的支撑集 ,φ 在 x_0 处可微, 那么我们有

$$\nabla_x c(x_0,y_0)=\nabla\phi(x_0).$$

这是由于 $\phi^c(y)=\inf\limits_{x\in X}c(x,y)-\phi(x)$, 因此最优情况下右侧梯度消失, 有:

$$\nabla_x c(x,y)=\nabla\phi(x)$$

因而, 最优传输映射为 $T^*(x)=(\nabla_x c(x,\cdot))^{-1}(\nabla\phi(x))$.

换用 *c-subdifferential* 语言来写, 有:

定理 ([2] 的 Theorem 1): 设 $\varphi(x)$ 为 c-concave 函数, 则 $\forall(x,y)\in\partial^c\varphi$ 有:

$$\nabla\varphi(x)=\nabla_x c(x,y)$$

由于当 $c(x,y)=h(x-y)$、h 为凸函数情形下, 最优传输映射与最优传输方案等价, 因此有: $\nabla h(x-T(x))=\nabla\phi(x)$, 由此可得:

$$T(x)=x-(\nabla h)^{-1}(\nabla\phi(x))$$

定理 ([1] 的 Theorem 1.17): 给定紧区域 $\Omega\subset\mathbb{R}^d$ 上的概率测度 μ 和 ν,μ 绝对连续 $\partial\Omega$ 零测度, 传输代价函数具有形式 $c(x,y)=h(x,y)$, 其中 h 是一个严格凸函数, 则存在唯一的最优传输方案 γ,γ 具有表示形式 $(\mathrm{id},T)_{\#}\mu$. 更进一步, 存在一个 Kantorovich 势能函数 φ,T 和 φ 通过下公式相联系 (最后一个等号是 Legende 对偶的性质)

$$T(x)=x-(\nabla h)^{-1}(\nabla\phi(x))=x-\nabla h^*(\nabla\phi(x))$$

若 $h=\frac{1}{2}|x|^2$,$\nabla h(x)=x$ 为恒等映射，即 $c(x,y)=\frac{1}{2}|x-y|^2$ 时有：

$$T(x)=x-\nabla\phi(x)=\nabla(\frac{1}{2}x^2-\phi)$$

称 $u_\phi(x)=\frac{\|x\|^2}{2}-\phi(x)$ 为 Brenier 势能函数, 即最优传输映射为 $T=\nabla u_\phi(x)$。这就是求解完对偶问题恢复出蒙日问题最优传输映射的方法!

另一方面，

$$\begin{aligned}\phi(x)&=\psi^{\bar{c}}(x)\\&=\inf_y\left[\frac{\|x-y\|^2}{2}-\psi(y)\right]\\&=\inf_y\left[\frac{\|x\|^2}{2}-x\cdot y+\frac{\|y\|^2}{2}-\psi(y)\right]\end{aligned}$$

此时

$$\begin{aligned}u_\phi(x)&=\frac{\|x\|^2}{2}-\phi(x)\\&=-\inf_y\left[-x\cdot y+\left(\frac{\|y\|^2}{2}-\psi(y)\right)\right]\\&=\sup_y\left[x\cdot y-\left(\frac{\|y\|^2}{2}-\psi(y)\right)\right]\end{aligned}$$

即 $u_\phi=(\frac{\|y\|^2}{2}-\psi(y))^*$, 为线性函数的上包络。因此，最优传输映射 $T=\nabla u$, 为一个凸函数的梯度!

定理 (Brenier): 给定紧区域 $\Omega\subset\mathbb{R}^a$ 上的概率测度 μ 和 ν,μ 绝对连续，$\partial\Omega$ 为 Lebesgue-零测度，传输代价函数为 $c(x,y)=\frac{1}{2}|x-y|^2$, 则存在唯一的最优传输方案 γ,γ 具有表示形式 $(\mathrm{id},T)_\#\mu$. 并且存在 Kantorovich 势 φ,T 和 φ 的关系为 $T(x)=x-\nabla\varphi(x)$. 更进一步，有 $T(x)=\nabla u(x)$, 其中 u 是一个凸函数，被称为 Brenier 势能函数.

Note: 最优传输映射的逆映射 T^{-1} 也是最优传输映射。由于 $T=\nabla u_\phi$, 因此 $T^{-1}=(\nabla u_\phi)^{-1}$. 由于互为 Legendre 对偶的两函数其微分互逆，因此 $T^{-1}=\nabla(u_\phi)^*$. 同前文 $\frac{\|x\|^2}{2}-\phi(x)$ 可知 $\frac{\|y\|^2}{2}-\psi(y)$ 为凸函数，因此 $T^{-1}=\nabla(\frac{\|y\|^2}{2}-\psi(y))$，其对应的 Brenier 势能函数为 $\frac{\|y\|^2}{2}-\psi(y)$,Kantorovoch 势能函数为 $\psi=\phi^c$。即最优传输映射与其逆映射 T^{-1} 对应的 Brenier 势能函数互为 Legendre 对偶，对应 Kantorovich 势能函数互为 c-transform。Legendre 对偶就是从函数的定义空间，与函数的切空间之间变换。最优传输映射 T 的 Brenier 势能函数为凸函数的线性包络（也就是凸函数的切（支撑）平面，其投影为 Power Diagram），那么其逆映射的 Brenier 势能函数为原凸函数（即原凸函数上的点（或切平面的对偶点形成的凸包），其投影为加权三角剖分）。

Note: 顾险峰老师关于最优传输几何观点的口诀：“代价变换支撑，支撑包络势能。势能微分映射，映射对偶凸形。”，相关的总结见其公众号文章最优传输几何观点的口诀,最优传输理论的全局几何观点.

2.3.2 Monge-Ampère 方程

给定 (Ω,μ) 和 (Ω^*,ν), , $\Omega,\Omega^*\subset\mathbb{R}^n$ 凸紧集，密度函数 $d\mu(x)=f(x)dx$, $d\nu(x)=g(y)dy$ 连续，Brenier 势能函数 $u:\Omega\to\mathbb{R}$ 是一个 C^2 光滑函数。最优传输映 $T:\Omega\to\Omega^*$ 保测度 $T_{\#}\mu=\nu$,$y=T(x)$, 因此我们得到 Jacobi 方程：

$$f(x)\,dx=g(y)\,dy\iff\frac{dy}{dx}=\frac{f(x)}{g(y)}\iff\det DT(x)=\frac{f(x)}{g\circ T(x)}$$

由 $T=\nabla u$, 我们得到 Monge-Ampère 方程：

$$\det D^2u(x)=\frac{f(x)}{g\circ\nabla u(x)}.$$

且具有第二类边界条件：

$$\nabla u(\Omega)=\Omega^*.$$

2.3.3 最差传输视角

本节研究 $c(x,y)=\frac{1}{2}|x-y|^2$ 下的最优传输的等价最差传输问题。从中体会 c-transform 与 Legendre 对偶的关系，并从另一视角推导出 Brenier 定理!

对于原始 Monge 问题：

$$\inf\left\{M(T):=\frac{1}{2}\int_X|x-T(x)|^2d\mu,\quad T_{\#}\mu=\nu\right\}$$

对目标函数 $M(T)$ 展开：

$$\frac{1}{2}\int_X|x-T(x)|^2d\mu=\frac{1}{2}\int_X|x|^2d\mu-\int_\Omega\langle x,T(x)\rangle d\mu+\frac{1}{2}\int_X|T(x)|^2d\mu$$

由于 T 保测度，第三项与 T 无关:

$$\frac{1}{2}\int_X|T(x)|^2d\mu=\frac{1}{2}\int_X|T(x)|^2f(x)dx=\frac{1}{2}\int_Y|y|^2g(y)dy=\frac{1}{2}\int_Y|y|^2d\nu$$

第一项亦与 T 无关。因此原始蒙日问题，可以转化为成本函数为 $<x,T(x)>$ 的最差传输问题：

$$\sup\left\{M(T):=\int_X\langle x,T(x)\rangle d\mu,\quad T_{\#}\mu=\nu\right\}$$

此时，对上述最差传输问题, 其对偶问题为：

$$\inf\left\{\int_X\tilde{\varphi}d\mu+\int_Y\tilde{\psi}d\nu:\tilde{\varphi}(x)+\tilde{\psi}(y)\geq\langle x,y\rangle\right\}$$

因此，此时的 c-transform 就变为 Legendre 变换。类似上文可以将 $(\tilde{\varphi},\tilde{\psi})$, 变至可行且更优的 $(\tilde{\psi}^*,\tilde{\psi})$, 再到可行且更优的 $(\tilde{\psi}^*,\tilde{\psi}^{**})$，因此对偶问题亦变为 (取 $\tilde{\varphi}=\tilde{\psi}^*$)：

$$\inf\left\{\int_X\tilde{\varphi}d\mu+\int_Y\tilde{\varphi}^*d\nu:\tilde{\varphi}(x)\ is\ convex\ function\rangle\right\}$$

其中 $\tilde{\varphi}$ 为某个凸函数 ψ 的 Legender 对偶，亦为凸函数。上述对偶问题最优时有 $\tilde{\varphi}(x)+\tilde{\varphi}^*(y)=\langle x,y\rangle$. 由于 $\tilde{\varphi}^*(y)=\sup_x < x,y > -\tilde{\varphi}(x)$。因此最优时，右侧梯度消失，且当前最差传输问题等价于原最优传输问题，最优传输映射存在。因此最优传输映射 $T(x)=\nabla\tilde{\varphi}$, 又验证了 Brenier 定理，即最优传输映射是某个凸函数的梯度。

那这个 $\tilde{\varphi}$ 和原问题的 φ 有啥关系？这可以从原问题的对偶问题得出。

原问题对偶的可行解满足的约束：$\varphi(x)+\psi(y)\le\frac{|x-y|^2}{2}$, 因此有：

$$x\cdot y\le\left[\frac{|x|^2}{2}-\varphi(x)\right]+\left[\frac{|y|^2}{2}-\psi(y)\right]$$

取 $\tilde{\varphi}=\frac{|x|^2}{2}-\varphi(x),\tilde{\psi}=\frac{|y|^2}{2}-\psi(x)$. 此时最优传输问题的对偶问题变为

$$M-\sup\left\{\int_X\tilde{\varphi}d\mu+\int_Y\tilde{\varphi}^*d\nu:\tilde{\varphi}(x)+\tilde{\varphi}^*(y)\ge\langle x,y\rangle\right\}$$

其中 $M=\int_X\frac{|x|^2}{2}d\mu+\int_Y\frac{|y|^2}{2}d\nu$ 为常数. 这也恢复到上面推导的最差传输对偶问题. 这样就知道 $\tilde{\varphi}=\frac{|x|^2}{2}-\varphi$. 这就和上一节的 Brenier 定理的形式完全一致。

对于上述 $T=\nabla\tilde{\varphi}$，$\tilde{\varphi}(x)+\tilde{\varphi}^*(y)=< x,y >$ 如果说其确为原 Monge 最优传输问题的解，还需要解决 $T=\nabla\tilde{\varphi}$ 保测度（可行性），以及最优性的问题。

最优性的证明相对简单，这里先假设上述 $T=\nabla\tilde{\varphi}$ 保测度，其中 $\tilde{\varphi}(x)+\tilde{\varphi}^*(y)=< x,y >$，$\tilde{\varphi}(x)$ 为对偶问题的最优解。

证明（最优性）： 任取保测映射 S,$S_\sharp\mu=\nu$. 其传输代价为 $J(S)=\int_X\frac{\|x-S(x)\|_2^2}{2}d\mu=\int_X\frac{|x|^2}{2}-< x,S(x) >+\frac{|S(x)|^2}{2}d\mu$.

根据对偶问题的约束条件有：$\tilde{\varphi}(x)+\tilde{\psi}(S(x))\ge< x,S(x) >$

因此 $J(S)\ge\int_X\frac{|x|^2}{2}-(\tilde{\varphi}(x)+\tilde{\psi}(S(x)))+\frac{|S(x)|^2}{2}d\mu=\int_X\frac{|x|^2}{2}-\tilde{\varphi}(x)d\mu+\int_Y\frac{|y|^2}{2}-\tilde{\psi}(y)d\nu$.

最后一个等式是基于 S 保测度。同样利用 T 的保测度，有 $J(S)\ge\int_x\frac{1}{2}|x|^2-\tilde{\varphi}(x)+\frac{1}{2}|T(x)|^2-\tilde{\psi}(T(x))d\mu$ 。此时利用 $\tilde{\psi}(T(x))+\tilde{\varphi}(x)=< x,T(x) >$, 可得不等号右侧正好为 $J(T)$. 因此有：$\forall S,S_\sharp\mu=\nu,J(S)\ge J(T)$ ■

下证可行性，即证 $\forall h\in C^0(\bar{Y})$, 有 $\int_X h(T(x))d\mu(x)=\int_Y h(y)d\nu(y)$. 下面的证明利用了变分思想。

证明（保测度）： 取 $\tilde{\psi}_\varepsilon(y)=\tilde{\psi}(y)+\varepsilon h(y),\tilde{\varphi}_\varepsilon(x)=\sup_{y\in Y}\{xy-\tilde{\psi}(y)-\varepsilon h(y)\}$. 由于 $\tilde{\psi}_\varepsilon(y)+\tilde{\varphi}_\varepsilon(x)\ge xy,\forall x\in X,y\in Y$, 因此 $(\tilde{\varphi}_\varepsilon(x),\tilde{\psi}_\varepsilon(y))$ 亦为对偶问题的可行解。而 $(\tilde{\varphi},\tilde{\psi})$ 为对偶问题的最优解，取 L 表示对偶问题的目标函数，因此有

$$\begin{aligned}0&\le\frac{L(\tilde{\varphi}_\varepsilon,\tilde{\psi}_\varepsilon)-L(\tilde{\varphi},\tilde{\psi})}{\varepsilon}\\&=\int_X\frac{\tilde{\varphi}_\varepsilon(x)-\tilde{\varphi}(x)}{\varepsilon}d\mu+\int_Y h(y)dv(y)\end{aligned}$$

且当 $\varepsilon \to 0$ 时, $\delta L = \lim_{\varepsilon\to 0} \frac{L(\tilde{\varphi}_\varepsilon, \tilde{\psi}_\varepsilon) - L(\tilde{\varphi}, \tilde{\psi})}{\varepsilon} = 0$. 即 $\int_X \frac{\tilde{\varphi}_\varepsilon(x) - \tilde{\varphi}(x)}{\varepsilon} d\mu + \int_Y h(y) dv(y) = 0$. 下只需证明当 $\varepsilon \to 0$ 时 $\int_X \frac{\tilde{\varphi}_\varepsilon(x) - \tilde{\varphi}(x)}{\varepsilon} d\mu = -\int_X h(T(x)) d\mu(x)$ 即可。

对于 $\tilde{\varphi}_\varepsilon(x) = \sup_{y\in Y}\{xy - \tilde{\psi}(y) - \varepsilon h(y)\}$, 当 $\varepsilon \to 0$, 其对应的最优的 $y_\varepsilon = T(x) + o(1)$. (相信其是对的)

一方面，由于 $\tilde{\varphi}(x) = \sup_{y\in Y}\{xy - \tilde{\psi}(y)\}$, 因此 $\tilde{\varphi}(x) \geq xy_\varepsilon - \tilde{\psi}(y_\varepsilon)$. 因此，$\tilde{\varphi}_\varepsilon(x) - \tilde{\varphi}(x) \leq -\varepsilon h(y_\varepsilon) = -\varepsilon h(T(x)) + o(\varepsilon)$.

另一方面, 由于 $\tilde{\varphi}_\varepsilon(x) = \sup_{y\in Y}\{xy - \tilde{\psi}(y) - \varepsilon h(y)\}$, 取 $y = T(x)$ 有 $\tilde{\varphi}_\varepsilon(x) \geq xT(x) - \tilde{\psi}(T(x)) - \varepsilon h(T(x))$. 而 $\varphi(x) = xT(x) - \varphi(T(x))$, 因此有:$\tilde{\varphi}_\varepsilon(x) - \tilde{\varphi}(x) \geq -\varepsilon h(T(x))$.

综合上述两个方面知, 当 $\varepsilon \to 0$ 时, $\tilde{\varphi}_\varepsilon(x) - \tilde{\varphi}(x) = -\varepsilon h(T(x))$。即 $\int_X \frac{\tilde{\varphi}_\varepsilon(x) - \tilde{\varphi}(x)}{\varepsilon} d\mu = -\int_X h(T(x)) d\mu$. 由此可得 $T = \nabla\tilde{\varphi}$ 满足保测度条件: $\forall h \in C^0(\bar{Y})$, 有 $\int_X h(T(x)) d\mu(x) = \int_Y h(y) d\nu(y)$ ■

2.4 A Formal Proof

2.5 Semi-discrete OT 与计算几何视角

半离散最优传输即对于源域 X, 目标域 Y 一个是连续空间另一个是离散空间。本篇笔记介绍目标域 Y 为离散空间的情形，相当于定点输运，应用场景比较多，如选址场景。此时，Kantorovich 问题的对偶问题:

$$\sup_{\psi\in\bar{c}-conc(Y)} \left[\int_X \psi^{\bar{c}}(x) d\mu + \int_Y \psi(y) d\nu\right]$$

由于 ν 为离散测度，$\nu = \frac{1}{n}\sum_{j=1}^n \nu_i \delta(y - y_j)$ 可改写为:

$$\sup_{\psi\in\bar{c}-conc(Y)} \int_X \psi^{\bar{c}}(x) u(x) dx + \sum_{j=1}^n \psi_j \nu_j$$

其中 $u(x)$ 为概率密度 $d\mu = u(x)dx$,ν_j 为 y_j 处的目标概率。

记上述优化问题的目标函数为 $F(\psi) = F(\psi_1, \psi_2, \ldots \psi_n)$，则

$$\begin{aligned}
F(\psi) = \quad & F(\psi_1, \psi_2, \ldots \psi_n) \\
= \quad & \int_X \psi^{\bar{c}}(x) u(x) dx + \sum_{j=1}^n \psi_j \nu_j \\
= \quad & \int_X \inf_{y_j\in Y} \left[c(x, y_j) - \psi_j\right] u(x) dx + \sum_{j=1}^n \psi_j \nu_j \\
= \quad & \sum_{j=1}^n \int_{\mathrm{L}ag_{\psi}^c(y_j)} \left(c(x, y_j) - \psi_j\right) u(x) dx + \sum_{j=1}^n \psi_j \nu_j
\end{aligned}$$

其中最后一个等号是对被积区域 X 也就是源域进行划分，由 $\psi^{\bar{c}}$ 的定义，划分正好就是 y_j 对应的 *Laguerre cell*：

$$\mathrm{L}ag_{\psi}^{c}(y_j) = \{x \in X \quad | \quad c(x, y_j) - \psi_j \leq c(x, y_k) - \psi_k, \forall k \neq j\}$$

特别的，当 $c(x-y) = \frac{1}{2}|x-y|^2$ 时，*Laguerre cell* 就是 *Power Diagram*(准确而言应该去掉那个 1/2，不过这个系数不影响最优传输方案的求取，总的传输代价也就差一个常数倍)。

. 本节的剩余内容将介绍 $F(\psi)$ 为凹函数的性质，并计算最优情况下 ψ^* 的次微分，即最优传输映射方案的支撑集，也就是最优传输映射。关于目标函数梯度与 Hessian 矩阵的推导以及求解半离散最优传输的数值方法将在 Sec.3.12中给出。

定理: 上述半离散最优传输问题的目标函数 $F(\psi)$ 为凹函数。

证明: 设 $\sigma : X \to \{1\ 2\ \cdots, n\}$ 为分配。

考察函数 $G(\sigma, [\psi_1, \ldots \psi_n]) = \int_X \left(c(x, y_{\sigma(x)}) - \psi_{\sigma(x)}\right) u(x)dx$

对 X 按照给定的分配进行划分，则

$$\begin{aligned} G(\sigma, \psi) &= \sum_j \int_{\sigma^{-1}(j)} \left(c(x, y_j) - \psi_j\right) u(x)dx \\ &= \underbrace{\sum_j \int_{\sigma^{-1}(j)} c(x, y_j)u(x)dx}_{\text{与 } \psi \text{ 无关}} - \sum_j \psi_j \int_{\sigma^{-1}(j)} u(x)dx. \end{aligned}$$

其中 $\sigma^{-1}(j) = \{x \in X | \sigma(x) = j\}$. 因此 G 是关于 ψ 的仿射变换。

可以发现: 给定 ψ, 若分配 σ 取上述 *Laguerre cell* 的分配，则 $G(\sigma, \psi)$ 取得最小.

这是因为 $G = \int_X \left(c(x, y_{\sigma(x)}) - \psi_{\sigma(x)}\right) u(x)dx \geq \int_X \inf_{y_j \in Y} \left[c(x, y_j) - \psi_j\right] u(x)dx$

即 $G = \sum_j \int_{\sigma^{-1}(j)} \left(c(x, y_j) - \psi_j\right) u(x)dx \geq \sum_j \int_{\mathrm{L}ag_{\psi}^{c}(y_j)} \left(c(x, y_j) - \psi_j\right) u(x)dx.$

记 *Laguerre cell* 对应的的分配为：$\sigma^{\psi}(x) = \arg\min_j \left[c(x, y_j) - \psi_j\right]$，则半离散最优传输为的目标函数变为：

$$\begin{aligned} F(\psi) &= G(\sigma^{\psi}, \psi) + \sum_{j=1}^{n} \psi_j \nu_j \\ &= \inf_{\sigma} G(\sigma, \psi) + \sum_{j=1}^{n} \psi_j \nu_j \end{aligned}$$

由于首项为一族仿射函数的的下确界，为凹函数；第二项为线性函数亦为凹函数。因此 $F(\psi)$ 为凹函数 ■.

半离散最优传输问题的对偶为 $\max_{\psi} F(\psi)$, 又 $F(\psi)$ 为凹函数，因此容易采用任意凸优化方法找到最优解!

下面来分析半离散传输方案的最优传输映射！也就等价于讨论最优情况下 ψ^* 的 c-次微分。

定理（半离散最优传输方案）： 设 $x \in X$, 且 $x \in \mathrm{L}ag_{\psi}^{c}(y_j)$，则半离散最优传输的最优传输方案将 $x \mapsto y_j$.

证明: 设 ψ 为最优的 Kantorovich 势能函数，则

$$\begin{aligned}
[\partial^c \psi](x) = & \quad \{y_k \mid \psi^c(x) + \psi_k = c(x, y_k)\} \\
= & \quad \left\{ y_k \middle| \inf_{y_l} [c(x, y_l) - \psi_l] + \psi_k = c(x, y_k) \right\} \\
= & \quad \{y_k \mid c(x, y_j) - \psi_j + \psi_k = c(x, y_k)\} \\
= & \quad \{y_k \mid c(x, y_j) - \psi_j = c(x, y_k) - \psi_k\} \\
= & \quad \{y_j\}
\end{aligned}$$

其中第三个等号是因为 $x \in \mathrm{L}ag_{\psi}^{c}(y_j)$ 的假定。 ■

半离散最优传输问题，每一个传输映射就是一个分配 σ. 对于 $Y = \{y_1, y_2, \cdots, y_n\}$ 和目标质量 $\{\nu_1, \nu_2, \cdots, \nu_n\}$，每给一个 ψ, 即可对求出每个 y_j 关联的 *Laguerre cell* 区域 $\mathrm{L}ag_{\psi}^{c}(y_j)$，此区域就关联一个传输映射！更进一步，根据 Sec.3.12推导的一阶最优条件——最优时 $v_j = \mathrm{L}ag_{\psi}^{c}(y_j), \forall j \in \{1, 2, \cdots, n\}$ 知，求解半离散最优传输问题的对偶问题本质上就是不断调整 ψ, 使得 y_j 对应的 *Laguerre cell* 区域上的质量正好为目标质量 ν_j. 最终的最优传输映射就是把任意 $x \in X$，映射到其所在 *Laguerre cell* 区域关联的 $y_j \in Y$.

下面，不加证明地给出 ψ 和 *Laguerre cell* 的关系：

定理: 设 $Y = \{y_1, y_2, \cdots, y_n\}$ 为大小为 n 的点集. 若 ψ 为 Y 上的任意函数，且 $\forall j, \mathrm{L}ag_{\psi}^{c}(y_j)$ 非空, 则 ψ 为 $\bar{c}$-concave 函数。

定理: 设 $Y = \{y_1, y_2, \cdots, y_n\}$ 为大小为 n 的点集. 若 ψ 为 Y 上的 $\bar{c}$-concave 函数，则其对应的 *Laguerre cell* 区域 $\mathrm{L}ag_{\psi}^{c}(y_j)$ 非空，$j \in \{1, 2, \cdots, n\}$。

上述两个定理的证明见文献[2] 的 Sec7.3 定理 3、4.

2.5.1 补充——Voronoi Diagram and Power Diagram

解释一些概念的区别与联系，Voronoi 图及其对偶 Delaunay triangulation，Power diagram 及其对偶 Weighted Delaunay triangulation(又称 Regular triangulation)，重心 Voronoi 图（CVT）(特殊的 Voronoi 图，每个胞腔的重心与对应点重合)，容积约束的 Voronoi 图（CCVT）(本质为一个最优传输问题)，约束 Voronoi(把 Voronoi 约束在某个 $\mathbb{R}^n$ 的子空间上，如球面)

学习资料:Voronoi 图与 Delaunay 三角剖分，Power Diagram 与 Regular triangulation，视频 1，视频 2.

2.6 流体力学下的 OT 与 Benamou-Brenier 定理

Brenier 于 2000 年左右提出了 OT 问题的在流体力学视角下的求解方法，也就是 Benamou-Brenier 定理，将最优传输等价于连续方程中动能最小的解，即与最小流等价。

具体而言，有如下事实：

- 求解损失函数为 $c(x,y)=|x-y|^p$ 形式的的最优传输问题，等价于在 Wasserstein Space 找链接源域与目标域的恒速测地线。这是因为从最优传输方案可以构造出恒速测地线，从测地线也可以重构出最优传输。
- 求恒速测地线，通过最小化 $\int_0^1 |\mu^{'}(t)|^p dt$, 其中 $\mu(t)$ 可以看做 $\mathbb{W}_p$ 空间下弧长参数化的曲线。
- 在 Wasserstein Space ($\mathbb{W}_p$) 下，$|\mu^{'}(t)|^p=\inf\limits_{v_t}\int_\Omega |v_t|^p d\mu_t$, 其中 v,μ 为满足连续方程的速度场和测度，尽管连续方程的解不唯一，但是在 L^p 范数意义下, 度量的导数 $|\mu^{'}(t)|$ 等于所有可能速度场的最小值。

2.6.1 流体力学相关背景知识

流体力学有两种视角，欧拉视角（Eulerian, 固定空间位置，测密度、速度场）、拉格朗日视角（Lagrangian，跟随粒子运动）。拉格朗日视角下把粒子的位置记为 $\gamma(t,x)$, 或 $\gamma_x(t)$ 用粒子的初始为啥对粒子做标记（虽然这个符号可能与偏导的符号混淆，但是很多参考书都是这么用的，所以这里就暂且保留）；欧拉视角下，密度、速度场是位置的函数，随时间 t 变化，分别记为 $\rho(t,x), v(t,x)$, 或简记为 $\rho_t(x), v_t(x)$，两者通过如下 ODE 联系：

$$\begin{cases} \gamma_x^{'}(t)=v_t(\gamma_x(t)) \\ \gamma_x(0)=x \end{cases}$$

上述带初值的 1 阶 ODE 定义了一个映射 $Y_t: x\mapsto \gamma_x(t)$, 表示粒子初始时刻经过时间 t 后的位置。将其视为一个传输（粒子）映射（又可称之为 *flow*），那么时间 t 下的密度（理解成测度）$\rho_t=(Y_t)_\sharp\rho_0$, 即 $(\forall A\subset\mathbb{R}^n, \rho_t(A)=\rho_0(Y_t^{-1}(A)))$。这个角度可将流体力学与（最优）传输关联，即给定速度场下的流与传输方案对应，特别的，最小流与

最优传输对应！流体运动过程中 v_t, ρ_t 满足流体的连续方程：

$$\partial_t \rho_t + \nabla \cdot (\rho_t v_t) = 0, \forall t \in (0,1)$$

给定速度场 v_t 后，上述 $\rho_t = (Y_t)_\sharp \rho_0$ 正是流体连续方程的弱解或分布意义下的解，并且是唯一解（定理及细节见参考文献 [1]）。因此，连续方程的解对应传输方案，其中总动能最小的解，对应最优传输。

连续方程简介:单位时间封闭曲面流出（含流入）的质量 $\oint \rho v ds \overset{Guass/Stokes}{=} \int \nabla \cdot (\rho v) dV$，单位时间此区域质量的变化 $-\frac{\partial}{\partial t}\int \rho dV$，两者属于两种角度，质量不能凭空产生或消失（只有流入何流出），因而有 $\int \nabla \cdot (\rho v) dV + \frac{\partial}{\partial t}\int \rho dV = 0$.

2.6.2 Benamou-Brenier 公式与恒速测地线

定理（Benamou-Brenier）： 令 $\mu_0, \mu_1 \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$，记 $\dot{X}_t = v_t(X_t), \|v_t\|^2_{\mu_t} := \int \|v_t\|^2 \mathrm{d}\mu_t$, 则

$$W_2^2(\mu_0,\mu_1) = \inf_{v_t,\mu_t} \left\{ \int_0^1 \|v_t\|^2_{\mu_t} \,\mathrm{d}t \,\middle|\, (\mu_t, v_t)_{t\in[0,1]} \, solves \, \partial_t \mu_t + \mathrm{div}(\mu_t v_t) = 0, \mu_t v_t \cdot \mathbf{n}|_{\partial\Omega} = 0 \right\}.$$

最优传输曲线 $(\mu_t)_{t\in[0,1]}$ 唯一，且 $X_t \sim \mu_t$。其中：$X_t = (1-t)X_0 + tX_1$, 且 $(X_0, X_1) \sim \bar{\gamma} \in \Pi[\mu_0,\mu_1]$，$\bar{\gamma}$ 为最优传输方案。

Note 在正式开始前先回忆一下柯西不等式 $< a, b > \leq ||a|| \cdot ||b||$, 等号成立当且仅当 a, b 同向共线；其积分形式为 $(\int_a^b fg dx)^2 \leq \int_a^b f^2 dx \int_a^b g^2 dx$, 取 $g \equiv 1$, 因而有 $(\int_0^1 f dx)^2 \leq \int_0^1 f^2 dx$, 等号成立当且仅当 $f(x) \equiv Const.$

证明： 通过如下等式以及不等式即可证明此定理。

$$\begin{aligned} W_2^2(\mu_0,\mu_1) &\leq \mathbb{E}[\|X_0 - X_1\|^2] = \mathbb{E}\Big[\Big\|\int_0^1 \dot{X}_t \mathrm{d}t\Big\|^2\Big] \\ &\leq \mathbb{E}\Big[\Big\|\int_0^1 \|\dot{X}_t\|^2 \mathrm{d}t\Big\|\Big] = \int_0^1 \mathbb{E}[\|\dot{X}_t\|^2]\mathrm{d}t = \int_0^1 \|v_t\|^2_{\mu_t}\mathrm{d}t. \end{aligned}$$

第一个不等号变成等号，当且仅当 $X_1 = T(X_0)$, 其中 T 为 $\mu_0 \to \mu_1$ 的最优传输映射，此时存在唯一的最优传输方案 $\gamma \in \Pi[\mu_0,\mu_1]$. 第二个不等式变成等式当且仅当 $\dot{X}_t \equiv Const, \forall t \in [0,1]$. 由边界条件知 $\dot{X}_t \equiv X_1 - X_0$, 由于 $X_t = X_0 + \int_0^t \dot{X}_t dt$, 因此有 $X_t = (1-t)X_0 + tX_1$. ■

由于 $X_t \sim \mu_t$，因此亦有 $\mu_t = [(1-t)id + tT]\sharp\mu_0, T$ 为上述 $\mu_0 \to \mu_1$ 的最优传输映射。$(\mu_t)_{t\in[0,1]}$ 在 $\mathcal{P}_2(\Omega)$ 上连接了 μ_0, μ_1，且为一种最小作用量，且 $\dot{X}_t \equiv Const$, 类似于 (恒速，也就是弧长参数化下) 测地线的概念。其确为 $\mathbb{W}_2$ 空间 (测度空间 $\mathcal{P}_2(\Omega)$, 与其上的 $\mathbb{W}_2$ 距离定义的度量空间) 下连接概率测度 μ_0, μ_1 的恒速测地线！

定义（$\mathbb{W}_p$ 空间下的恒速测地线）： 令 $\mu_0,\mu_1\in\mathcal{P}_p(\mathbb{R}^d)$,T 为 $\mu_0\to\mu_1$ 的最优传输映射，则 $(\mu_t)_{t\in[0,1]}$ 为 $\mathcal{P}_p(\mathbb{R}^d)$ 上连接 μ_0,μ_1 的恒速（弧长参数化下）测地线：

$$\mu_t=[(1-t)\mathrm{id}+tT]_{\#}\mu_0$$

特别的，$p=2$ 时，由 Brenier 定理有：$T=id-\nabla\phi$,ϕ 为 Kantorovich 势能函数。则 $\mathscr{P}_2(\Omega)$ 上的恒速测地线为 $\mu_t=[\mathrm{id}+t\nabla\phi]_{\#}\mu_0$. 此外，测地线都是恒速的（恒速率，不代表没有加速度即不代表是恒速度）！弧长参数化后，一阶微分 $|\dot\gamma|\equiv1$，因此

$$\frac{d}{dt}|\dot\gamma|^2=2\langle\dot\gamma,\ddot\gamma\rangle=0$$

，即恒速率，但是 $\ddot\gamma$ 可存在法向分量，即切空间上对应的测地曲率为 0，法方向上对应的法曲率允许存在。

因此，若 $(\mu_t)_{t\in[0,1]}$ 为 $\mathbb{W}_p$ 空间下的恒速测地线，很自然的有：

$$W_p(\mu_t,\mu_s)=|t-s|W_p(\mu_0,\mu_1)$$

这个式子也可以当作 W_p 空间上恒速测地线的定义。

当然上述定义都是假设最优传输映射 T 存在。如果不存在，测地线得用最优传输方案 γ 表示，此时测地线 $\mu_t=(\pi_t)_\sharp\gamma$, 其中 $\pi_t(x,y)=(1-t)x+ty$, $(\pi_0)_\sharp\gamma=\mu_0,(\pi_1)_\sharp\gamma=\mu_1$. 若最优传输方案方案不唯一则测地线也不唯一。换言之，每个最优传输方案诱导了一条 W_p 空间下的测地线。

延用这里的记号，证明上述等式成立的方式为：

证明： 令 $\gamma_{s,t}:=(\pi_s,\pi_t)_{\#}\gamma\in\Gamma(\mu_s,\mu_t)$，则：

$$\begin{aligned}W_p(\mu_s,\mu_t)&\le\left(\int_{X\times X}|z-z'|^pd\gamma_{s,t}(z,z')\right)^{\frac1p}\\&=\left(\int_{X\times X}|\pi_s(x,y)-\pi_t(x,y)|^pd\gamma(x,y)\right)^{\frac1p}\\&=|t-s|\left(\int_{X\times X}|x-y|^pd\gamma\right)^{\frac1p}=|t-s|W_p(\mu_0,\mu_1)\end{aligned}$$

对 $[0,1]$ 分 $[0,s],[s,t],[t,1]$，应用上述不等式有：

$$W_p(\mu_0,\mu_s)+W_p(\mu_s,\mu_t)+W_p(\mu_t,\mu_1)\le(s+(t-s)+1-t)\,W_p(\mu_0,\mu_1)=W_p(\mu_0,\mu_1).$$

同时由于 W_p 为距离，有三角不等式，因而另一方向成立。

因此 $W_p(\mu_t,\mu_s)=|t-s|W_p(\mu_0,\mu_1),\forall0\le s,t\le1.$ ■

Remark1: Brenamou-Brenier 公式为一泛函求极小问题。按照变分学里的记号，可定义 Action：

$$A[\mu_0,\mu_1]=\int_0^1 \|v_t\|_{\mu_t}^2\,\mathrm{d}t$$

因此，

$$W_2^2(\mu_0,\mu_1)=\inf_{v_t,\nu_t}\left\{A[\mu_0,\mu_1]\,\middle|\,(\mu_t,v_t)_{t\in[0,1]}\ solves\ \partial_t\mu_t+\mathrm{div}(\mu_t v_t)=0,\mu_t v_t\cdot\mathbf{n}|_{\partial\Omega}=0\right\}$$

Remark2: 一般二维流形上定义的黎曼（测地）距离：

$$d_M(x,y):=\inf\left\{\int_a^b|\dot\gamma(t)|dt\,\middle|\,\gamma{:}[a,b]\to M,\gamma(a)=x,\gamma(b)=y\right\}$$

考虑到曲线都可以进行弧长参数化。进一步可以证明黎曼距离有如下等价定义：

$$d_M^2(x,y)=\inf\left\{\int_0^1|\dot\gamma(t)|^2dt\,\middle|\,\gamma\colon[0,1]\to M,\gamma(0)=x,\gamma(1)=y\right\}$$

W_p 距离相当于 $\mathbb{W}_p$ 空间上两点的测地距离。类比可以定义 $\mathbb{W}_p$ 空间某点 μ_t 切平面上向量 $\dot\mu_t$ 的范数为：

$$\|\partial_t\mu_t\|_{\mu_t}^2:=\inf_{v_t}\left\{\int_\Omega|v_t|^2\mu_t dx\,\middle|\,\mathrm{div}(v_t\mu_t)=-\partial_t\mu_t,\mu_t v_t\cdot\mathbf{n}|_{\partial\Omega}=0\right\}$$

则 W_p 可写成如下形式：$W_2^2(\bar\mu_0,\bar\mu_1)=\inf_{\mu_t}\left\{\int_0^1\|\partial_t\mu_t\|_{\mu_t}^2\,dt\,\middle|\,\mu_0=\bar\mu_0,\mu_1=\bar\mu_1\right\}$

根据 Benamou-Brenier 定理，时变密度场 $\rho_t(x)$ 从流体初始时刻 $t=0$ 的 ρ_0，在初始速度场 $v_0(x)$ 作用下，受流体连续方程的约束，变到目标密度场即 $t=1$ 下的 ρ_1. 在所有可能的“运动过程”（Eulerian 视角下，满足连续方程的所有可能的时变 $\rho_t(x),v_t(x)$）中，总能量（动能）最小为 $\mathcal{W}_2^2(\rho_0,\rho_1)$，即 Benamou-Brenier 公式:

$$\mathcal{W}_2^2(\rho_0,\rho_1)=\min_{v(x,t),\rho(x,t)}\frac{1}{2}\int_0^1\int_{\mathbb{R}^n}\rho(x,t)\|v(x,t)\|_2^2dxdt$$

$$s.t.\quad\begin{cases}\rho(x,0)=\rho_0(x),\forall x\in\mathbb{R}^n\\ \rho(x,1)=\rho_1(x),\forall x\in\mathbb{R}^n\\ \frac{\partial}{\partial t}\rho(x,t)+\nabla\cdot(\rho(x,t)v(x,t))=0,\forall x\in\mathbb{R}^2,t\in(0,1)\end{cases}$$

Remark: 这里 ρ_t 表示的是密度，很多书都是用测度 α_t 代替，这就导致看起来似乎形式不一样，本质上是一样的，如果视 $d\alpha_t(x)$ 为 $\rho_t(x)dx$.

上述优化问题是一个非凸优化问题: 对于变量 (ρ,v)，函数 $f=\rho|v|^p$ 不是凸函数（Hessian 矩阵的行列式小于 0，非正定）。且约束非线性，因为有变量相乘的项，即多项式项：$\rho(x,t)v(x,t)$。通过引入流量 $J(x,t)=\rho(x,t)v(x,t)$，可将变量变为 $(\rho(x,t),J(x,t))$

变为一个凸优化问题, 且约束线性:

$$\mathcal{W}_2^2(\rho_0,\rho_1)=\min_{J(x,t),\rho(x,t)}\frac{1}{2}\int_0^1\int_{\mathbb{R}^n}\frac{\|J(x,t)\|_2^2}{\rho(x,t)}dxdt$$
$$s.t.\quad\begin{cases}\rho(x,0)=\rho_0(x),\forall x\in\mathbb{R}^n\\ \rho(x,1)=\rho_1(x),\forall x\in\mathbb{R}^n\\ \frac{\partial}{\partial_t}\rho(x,t)+\nabla\cdot J(x,t)=0,\forall x\in\mathbb{R}^n,t\in(0,1)\end{cases}$$

最优传输问题的流体力学视角在插值问题上有特殊的应用，因为它可视为给定初始状态和目标状态的“流动”过程，取不同时刻 $t\in(0,1)$ 下，流场的密度场就相当于插值点，相比于简单的凸组合插值更有几何意义（最优传输对应的流动过程是一条测地线，插值点为测地线上的点，而黎曼几何中测地线就类比欧式几何下的直线)！特别地，对于一个给定区域上的概率空间，两个概率分布间的最优传输为，概率空间所在流形上，连接两个点（概率分布）的测地线。

Note: 上述插值又称位移插值（*displacement interpolation,*）也称 *McCann* 插值。

上述仅为损失函数 $c(x,y)=|x-y|^p$，在 p=2 时对应的 $\mathcal{W}_2$ 的 *Eulerian* 形式，p 取其他值有不同的 *Eulerian* 形式。如 $p=1$,$\mathcal{W}_1(\rho_0,\rho_1)$ 为

$$\mathcal{W}_1(\rho_0,\rho_1)=\min_{J(x,t)}\int_{\mathbb{R}^n}\|J(x,t)\|_2dx$$
$$s.t.\quad\nabla\cdot J(x,t)=\rho_1(x)-\rho_0(x)$$

这个问题叫做 *Beckmann problem*, 也是一个凸优化问题，它与时间无关了，更方便计算；但在插值任务上，退化为普通的初始与目标状态间的凸组合。

2.6.3 Beckmann's minimal flow 问题

如果说 $p=2$ 下的 $\mathcal{W}_2$ 距离的 Eulerian 表示是时变、动态的，由流体运动的连续性方程约束；那么 p=1 下的 W_1 具体可以理解成一种静态的，或者理解成一段时间下的平均，由散度调节约束。Beckmann 问题最早是由 Beckmann 在 1950s 年代提出的，它与 Kantorovich 问题的提出是同时期的，但并不了解彼此的工作。但是，在损失函数 $c(x-y)=|x-y|$ 下，两者等价。下面阐述这件事。

在正式开始之前，我们先回顾 Kantorovich 问题的对偶问题（DP）：

$$(DP):\quad\max_{\substack{\phi(x),\psi(y)\\ \phi\oplus\psi\le c(x,y)}}\int_X\phi(x)d\mu+\int_Y\psi(y)d\nu=\max_{\phi\in c-conx(X)}\int_X\phi d\mu+\int_Y\phi^c d\nu$$

其中 $\phi^c(y)=\inf_{x\in X}[c(x,y)-\phi(x)]$，$c-conx(X):=\{f|\exists\xi(y)s.t.,f=\xi^{\bar{c}}\}$.

命题：若 $c:X\times X\to\mathbb{R}$ 为一个距离，则函数 $u:X\to\mathbb{R}$ 是 c-concave 当且仅当函数

$u(x)$ 是相对于距离 c 是 Lipschitz 连续，且 Lipschitz 常数小于 1。记 Lip_1 代表这类函数，则 $\forall u \in Lip_1$, 有 $u^c = -u$. 特别的取 $c = |x-y|$, 那前面的相对于距离 c 的 Lipschitz 就是常见的 Lipschitz.

证明："⇒",$u(x)$ 为 c-concave 函数，则 $\exists \chi(y), s.t., u = \chi^{\bar{c}} = \inf_{y\in X} c(x,y) - \chi(y)$. 因为映射 $x \mapsto c(x,y) - \chi(y)$ 是 Lip_1 函数（$|(c(x_1,y)-\chi(y)) - (c(x_2,y)-\chi(y))| = |c(x_1,y)-c(x_2,y)| \leq c(x_1,x_2)$（距离的三角不等式））. 因此 $\inf_{y\in X} c(x,y) - \chi(y)$ 也是 Lip_1 函数（一组有相同 Lipschitz 常量的 Lipschitz 函数的下确界有共同的连续模），因而 $u(x) \in Lip_1$.

"⇐", $\forall u \in Lip_1$, 有 $\inf_{y\in X} c(x,y) + u(y) \leq c(x,x) + u(x) = u(x)$(c 为距离). 另一方面，$u(x) \in Lip_1$（相对于距离 c），则 $u(x) - u(y) \leq c(x,y)$，则 $u(x) = c(x,y) + u(y)$，因而有 $u(x) \leq \inf_{y\in X} c(x,y) + u(y)$，综合有 $u(x) = \inf_{y\in X} c(x,y) + u(y)$，即得 $u(x)$ 为 c-concave 函数，且 $u = (-u)^{\bar{c}}$. 对等式换元，且利用距离 c 的对称性有 $u(y) = \inf_{x\in X} c(x,y) + u(x)$，因而 $u = (-u)^c$

由于 $u \in Lip_1, \Rightarrow -u \in Lip_1$, 则，$-u = (-(-u))^c$，即 $u = -u^c$ ■

现在再看 (DP) 问题有：

$$\max_{\phi\in c-conx(X)} \int_X \phi d\mu + \int_Y \phi^c d\nu = \max_{\phi\in c-conx(X)} \int_X \phi d(\mu-\nu) = \max_{\phi\in Lip_1} \int_X \phi d(\mu-\nu)$$

因此有

$$\mathcal{W}_1(\mu,\nu) = \max_{\phi\in Lip_1} \int_X \phi d(\mu-\nu)$$

这正是 Wasserstein GAN 那篇文章的损失函数。

现在，开始解决 Beckmann 问题，由于 Beckmann 问题亦是凸优化问题，下面推导 Beckmann 问题的对偶以说明其与 Kantorovich 问题的等价性。其定义为：

$$\min\left\{\int |\mathbf{w}(x)|\mathrm{d}x : \mathbf{w} : \Omega \to \mathbb{R}^d, \nabla\cdot\mathbf{w} = \mu - \nu\right\}$$

其中 $\mathbf{w}$ 为流量之意。更一般的表述为被积函数为 $H(\mathbf{w})$,H 为凸函数。其中散度约束条件在零流边界条件下（*no-flux boundary conditions*，边界处单位面积的法向流量为 0）的弱解意义下为：$-\int \nabla\phi\cdot\mathrm{d}\mathbf{w} = \int \phi \mathrm{d}(\mu-\nu), \forall \phi \in C^1(\overline{\Omega})$（也就是用测试函数再加上分部积分）。此时有：

$$\sup_\phi \int_\Omega \nabla\phi\cdot\mathbf{w}\mathrm{d}x + \int_\Omega \phi\mathrm{d}(\mu-\nu) = \begin{cases} 0 & \text{if } \nabla\cdot\mathbf{w} = \mu-\nu \\ +\infty & \text{otherwise.} \end{cases}$$

因此原 Beckmann 问题等价为：

$$\inf_{\mathbf{w}}\left(\int_\Omega |\mathbf{w}|\mathrm{d}x + \sup_\phi \int_\Omega \nabla\phi\cdot\mathbf{w}\mathrm{d}x + \int_\Omega \phi\mathrm{d}(\mu-\nu)\right)$$

交换 $\inf, \sup$ 得到其对偶问题：

$$\sup_{\phi}\left(\int_{\Omega}\phi \mathrm{d}(\mu-\nu)+\inf_{\mathbf{w}}\int_{\Omega}(|\mathbf{w}|-\nabla\phi\cdot\mathbf{w})\mathrm{d}x\right)$$

由于：

$$\inf_{\mathbf{w}}\int_{\Omega}(|\mathbf{w}|-\nabla\phi\cdot\mathbf{w})\mathrm{d}x=\begin{cases}0 & \text{if } |\nabla\phi|\leq 1\\ -\infty & \text{otherwise}\end{cases}$$

因此对偶问题等价于：

$$\sup\left\{\int_{\Omega}\phi \mathrm{d}(\mu-\nu) \quad : \quad |\nabla\phi|\leq 1\right\}$$

这正好与 $p=1$ 时 Wasserstein 对偶一致。由于两个问题都是凸问题（目标函数凸函数，约束线性），因此对偶问题与原问题等价。因此 Beckmann 问题等价与 $p=1$ 时 Wasserstein 问题。

2.7 Some Inequality

3 经典 OT 的求解

Kantorvich-OT 问题就是线性规划问题，可以用单纯形法（如网络单纯性法）以及其他的凸优化方法（如内点法）去计算，但复杂度高。因此添加一些正则化项，改变目标函数的性质或者解的结构对加快计算有利。本章仅围绕一种广泛采用的熵正则方法进行介绍，该正则化增加了目标函数的凸性，使得更利于采用凸优化方法进行求解。，本章只是笔记形式，更多关于算法的细节及实现，请参考Peyré、顾险峰等人的书、POT 库，以及一些学者（如清华吴昊（男，利用/引入一些代数结构，加快计算）、人大许洪腾等）人的论文。

低秩约束正则化为近期产生的另一种正则化，这里不详细介绍（类似于奇异值分解，当期望传输方案的秩比较小的时候，把待求传输方案进行低秩分解为 $U\Sigma_r V$ 的形式，其中 Σ_r 为 r 维对角阵,r 小于传输方案的维度；然后基于 OT 目标函数是 **Tr** 运算，利用迹的性质可以分别把新产生的三个变量当作新的传输方案，利用轮换坐标下降产生三个新的最优传输问题进行求解）。详细请参考：Low-Rank Sinkhorn Factorization, 更多相关工作见Meyer Scetbon的主页。

3.1 一维最优传输的解析解

在正式开始最优传输的通用解法前，先看最简单的 1d 情形, 即 $X \subset \mathbb{R}, Y \subset \mathbb{R}$，它有解析形式。直观上，对于离散形式的一维最优传输问题，当源域与目标域大小相同时，最优传输映射就是把源域、目标域排好序，然后把 $X_{(i)}$ 映射到 $Y_{(i)}$. 其中 $X_{(i)}, Y_{(i)}$ 为次序统计量。对应的 Wasserstein 距离变为:

$$\mathcal{W}_p = \left[\sum_{i=1}^{n} \|X_{(i)} - Y_{(i)}\|^p \right]^{\frac{1}{p}}$$

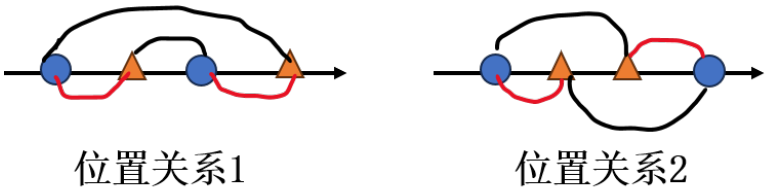


图 1: $\{X_{(i)}, X_{(j)}\}$, 与 $\{Y_{(i)}, Y_{(j)}\}$ 的两种位置关系

现证明为啥最优传输映射是 $T : X \to Y, X_{(i)} \mapsto Y_{(i)}$. 不妨设 $i < j$, 那么 $\{X_{(i)}, X_{(j)}\}$, 与 $\{Y_{(i)}, Y_{(j)}\}$ 无外乎图1所示的两种位置关系。不论如何都有 $\|X_{(j)} - Y_{(i)}\| + \|X_{(i)} - Y_{(j)}\| \geq \|X_{(i)} - Y_{(i)}\| + \|X_{(j)} - Y_{(j)}\|$, 因此最小化传输成本必然要求 $X_{(i)} \mapsto Y_{(i)}, \forall i \in \{1, 2, , \cdots, |X|\}$.

将其推广到连续空间上（直观上，非严谨）。上述过程做了件啥事? 就是从小到大排好顺序，按照顺序依次对应上。现在是连续了，那就从最小的开始，也就是从 $-\infty$ 开始，离散变连续，从一个小区间 $[-\infty, x]$,$x \to -\infty$ 开始传，传到对面的最小的某个小区间 $[-\infty, y]$，$y \to -\infty$，并且由有测度条件，有 $\nu([-\infty, y]) = \mu([-\infty, x])$, 即 $F_\nu(y) = F_\mu(x)$ 可得 $y = F_\nu^{-1}(F_\mu(x))$. 接下来，传小区间 $[x, x+\delta x]$, 将会传到小区间 $[y, y+\delta y]$, 并且要求 $\nu([y, y+\delta y]) = \mu([x, x+\delta x])$, 结合上个小区间的测度等式, 仍有 $\nu([-\infty, y]) = \mu([-\infty, x])$, 即 $F_\nu(y) = F_\mu(x)$. 以此类推，可得 $\forall x \in X \subset \mathbb{R}$, 最优传输映射 $T(x) = F_\nu^{-1}(F_\mu(x))$, 并且 Wasserstein 距离变为:

$$W_p^p = \int_X d^p(x, T(x)) d\mu = \int_X d^p(x, F_\nu^{-1}(F_\mu(x)) d\mu$$
$$\xrightarrow[dz = d\mu = f(x)dx]{z = F_\mu(x)} \int_0^1 d^p(F_\mu^{-1}(z), F_\nu^{-1}(z)) dz$$

当然了，累计分布 $F(x)$ 不一定有逆映射，比如在其非严格单增的地方，很多个相同的 x 映到相同的 y. 因而这里的逆定义为 $F^{[-1]}(x) := \inf\{t \in \mathbb{R} \mid F(t) \geq x\}$.

3.1.1 补充——c-循环单调与排序不等式

通过上文分析（如图1），我们很自然引出最优传输映射的性质——c 循环单调性。

定义（c-循环单调性): 设成本函数 $c : X \times Y \to \mathbb{R} \cup \{+\infty\}$, 称集合 $\Gamma \subset X \times Y$, 如果对任意 $k \in \mathbb{N}$, 任意有限点集 $(x_1, y_1), (x_2, y_2), \cdots, (x_k, y_k) \in \Gamma$, 都有：

$$\sum_{i=1}^{k} c(x_i, y_i) \leq \sum_{i=1}^{k} c(x_{\sigma(i)}, y_\sigma(i)), \forall \sigma \in S_k.$$

其中 S_k 为置换群。

我们不加证明的给出如下两个定理，反应最优传输映射与 c 循环单调的联系。下属定理之证明见文献 [1] 的 Page29-35。

定理: 若 γ 为关联连续的成本函数 c 的最优传输方案，则 $supp(\gamma)$ 是 c-循环单调的。

若成本函数 c 非连续，则该定理不一定成立。反之，基本上也成立。

定理（Rockafellar）： 给定传输成本函数 $c : X \times Y \to \mathbb{R}$(c 这里不取 $+\infty$), 若 $\emptyset \neq \Gamma \subset X \times Y$ 是 c-循环单调的, 则存在 c-凹函数 $\phi : X \to \mathbb{R} \cup \{-\infty\}$, 使得 $\Gamma \subset \partial^c \phi$,i.e.,

$$\Gamma \subset \{(x, y) \in X \times Y : \phi(x) + \phi^c(y) = c(x, y)\}$$

引入 c-循环单调的一个用途，就是证明最优传输 Kantorovich 问题与其对偶问题的等价性：$\min(KP) = \max(DP)$。

即如下定理：

定理（连续代价下 $\min(KP) = \max(DP)$）： 设成本函数 $c : X \times Y \to \mathbb{R}$ 一致连续且有界，X, Y 为 Polish 空间。则对偶问题存在一个解 (ϕ, ϕ^c)，且有 $\min(KP) = \max(DP)$ 成立。

上述定理可以推广至下半连续成本函数上：

定理（下半连续代价下 $\min(KP) = \max(DP)$）： 设 X, Y 为 Polish 空间，成本函数 $c : X \times Y \to \mathbb{R} \cup \{+\infty\}$ 下半连续且有界, 则 $\min(KP) = \max(DP)$.

同样，下半连续成本函数关联的最优传输方案是 c-循环单调的。

定理: 若传输代价 c 下半连续 (l.s.c)，γ 为最优传输方案，则 γ 集中于一个 c-循环单调的集合 Γ 上（通常 Γ 非闭集）。

最后，引入一个与 $c-$ 循环单调类似的排序不等式。

排序不等式: 给定 $a_1 \leq a_2 \leq \cdots \leq a_n$,$b_1 \leq b_2 \leq \cdots \leq b_n$,$\sigma \in S_n$ 为任一置换，并设 $r \in S_n, r\ i \mapsto n - i, i \in [1, 2, \cdots, n]$. 则

$$\sum_{i=1}^{n} a_i b_{r(i)} \leq \sum_{i=1}^{n} a_i b_{\sigma(i)} \leq \sum_{i=1}^{n} a_i b_i$$

换言之，顺序和 $\geq$ 乱序和 $\geq$ 逆序和。

从排序不等式也可窥探最优传输（一维）。成本函数为 $c = \frac{|x-y|^2}{2}$ 的最优传输等价于成本函数为 $c = xy$ 的最差传输。而顺序和的总成本将最高，因此，最优传输映射就是”顺序” 映射!

排序不等式参考资料：排序不等式 1，排序不等式与循环单调。

3.1.2 补充——一维随机变量生成的逆方法

一元最优传输有解析解，与其累计分布函数（c.d.f）有关。另一个累计分布有关的当属一维采样。

命题： 设 X 为一随机变量，$X \sim F$. 累计分布函数 F 的广义逆设为 $F^{-1}(u) = \inf\{x : F(x) \geq u\}$，则 $Y \sim F$，其中 $Y = F^{-1}(U), U$ 服从均匀分布 $\mathcal{U}[0,1]$。若 F 连续，则 $F(X) \sim \mathcal{U}[0,1]$.

证明： 我们知道累计分布函数 F 为递增、右连续、存在左极限的函数。我们将得到:

- F^- 单增不降
- $F(F^{-1}(u)) \geq u$, 若 F 连续，则 $F(F^{-1}(u)) = u$
- $F^{-1}(u) \leq x \Leftrightarrow u \leq F(x)$

则 $F_Y(y) = P(Y \leq y) = P(F^{-1}(U) \leq y) = P(U \leq F(y)) = F(y)$ 最后一个等式应用了 $U \sim \mathcal{U}[0,1]$. 即证 $Y = F^{-1}(U)$ 与 X 同分布。

命题的最后一部分 $F(X) = F(F^{-1}(U)) = U$, 最后一个等号利用 F 的连续性假设。

■

上述命题给出了采样随机变量 X 的方法：均匀采样 $U \sim \mathcal{U}[0,1]$, 则 $F^{-1}U$ 就是想要得到的样本。显然这需要知道分位函数 F^{-1}.

最后,我们给出一个经典的 Monte Carlo Markov Chain(MCMC)方法——Metropolis-Hastings Samploing Algorithm.

其思想为：构建一 Markov 随机过程，使得其极限分布为目标采样的分布 π. 这样，我们就可以运行这个随机过程一段时间 (*Burn-in*)，等它收敛以后，运行后获得的每个点都是目标分布的采样点。现在问题变为如何构建这个随机过程。

假设该 Markov 随机过程的一次概率转移矩阵为 $P = (P_{ij})$,P_{ij} 表示从状态 i 下一次移动到状态 j 的概率,i.e.,$P_{ij} = P(z_j|z_i)$. 我们知道其极限概率应满足 $\pi = P\pi.\pi$ 现在就是我们想要采样的概率分布。

我们引入上式的一个充分条件，称为 *Detailed Balance*:

$$\pi_i P_{ij} = \pi_j P_{ji}, \forall i, j$$

可以理解为平衡时从 i 流向 j 等于从 j 流回 i. 之所以说为充分条件，是因为: $\pi_j = \sum_i \pi_j P_{ji} = \sum_i \pi_i P_{i,j}$ 即得 $\pi = P\pi$.

因此，我们可以构建一个满足上述细致平衡方程的概率转移矩阵，这样其关联的 Markov 随机过程将满足极限方程: $\pi = P\pi$，π 为待采样分布。而构建上述概率转移矩阵不可谓不天才!

任取一个概率转移矩阵 P(称为 *proposal*), 做如下修正:

$$\widetilde{P}_{ij} = P_{ij} \cdot \min\{1, \frac{\pi_j P_{ji}}{\pi_i P_{ij}}\}$$

则 $\tilde{P}$ 基本上（由公式得行和一般会小于 1，不是概率转移矩阵）就是我们要找的概率转移矩阵。

你可以验证上述概率转移矩阵确实满足细致平衡方程:

$$\pi_i \widetilde{P}_{ij} = \begin{cases} \pi_i P_{ij} & \pi_j P_{ji} \geq \pi_i P_{ij} \\ \pi_j P_{ji}, & \pi_j P_{ji} < \pi_i P_{ij} \end{cases}$$

$$\pi_j \widetilde{P}_{ji} = \begin{cases} \pi_i P_{ij} & \pi_j P_{ji} \geq \pi_i P_{ij} \\ \pi_j P_{ji}, & \pi_j P_{ji} < \pi_i P_{ij} \end{cases}$$

因而 $\pi_i \widetilde{P}_{ij} = \pi_j \widetilde{P}_{ji}, \forall i, j$.

现在就把 $\tilde{P}$ 进行 normlize 一下就可以了: 将上述修正方程视为一种接受-拒绝策略，取 $U \sim \mathcal{U}[0,1]$, 若 $U \leq \min\{1, \frac{\pi_j P_{ji}}{\pi_i P_{ij}}\}$ 则允许 i 转移到 j, 否则不允许转移。算法见图2.

3.1.3 补充——多元高斯间的 Wasserstein 距离

对于两个多元高斯分布 $\alpha = \mathcal{N}(\boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma}_\alpha)$,$\beta = \mathcal{N}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)$, 其间的最优传输映射为:

$$T : x \mapsto \boldsymbol{\mu}_\beta + A(x - \boldsymbol{\mu}_\alpha)$$

其中，$A = \Sigma_\alpha^{-\frac{1}{2}} \left(\Sigma_\alpha^{\frac{1}{2}} \Sigma_\beta \Sigma_\alpha^{\frac{1}{2}}\right)^{\frac{1}{2}} \Sigma_\alpha^{-\frac{1}{2}} = A^{\mathrm{T}}$. 用一维来看，就是在调整其标准差或方差。

Wasserstein 距离为:

$$\mathcal{W}_2^2(\alpha, \beta) = \|\boldsymbol{\mu}_\alpha - \boldsymbol{\mu}_\beta\|^2 + \mathcal{B}(\boldsymbol{\Sigma}_\alpha, \boldsymbol{\Sigma}_\beta)^2$$

其中，$\mathcal{B}(\boldsymbol{\Sigma}_\alpha, \boldsymbol{\Sigma}_\beta)^2 \overset{\text{def.}}{=} \mathrm{Tr}\left(\boldsymbol{\Sigma}_\alpha + \boldsymbol{\Sigma}_\beta - 2(\boldsymbol{\Sigma}_\alpha^{1/2} \boldsymbol{\Sigma}_\beta \boldsymbol{\Sigma}_\alpha^{1/2})^{1/2}\right)$.

1. Initialize starting state $x^{(0)}$ and set $t = 0$

2. Burn-in: while samples have not converged, draw samples...

   - $x = x^{(t)}$
   - $t = t + 1$
   - sample $x^* \sim Q(x^*|x)$ # draw from proposal
   - sample $u \sim Uniform(0, 1)$ # draw acceptance threshold
     - if $u < A(x^*|x) = min\left(1, \frac{P(x^*)Q(x|x^*)}{P(x)Q(x^*|x)}\right)$, then $x^t = x^*$ # transition
     - else $x^t = x$ # stay in current state
   - Take samples from $P(x)$: reset $t = 0$, for $\mathbf{t} = 1 : N$
     - $x(t + 1) \leftarrow$ Draw sample $x(t)$

图 2: Metropolis-Hastings Samploing Algorithm

3.1.4 补充——概率空间或标准单纯形下的优化

手写笔记

3.2 Sinkhorn Algorithm

本节和下一小节介绍由法国学派发展了一套形式简洁、计算效率高、方便并行的熵正则方法，即在原目标函数中增加一正则项，且正则项为一种熵。本小节将熵函数作为正则项，而下一小节将相对熵作为正则项。需要指出，下一小节的熵正则本质上为一种近端点法，它优化的目标仍为原目标函数，只是每个迭代步上加上由相对熵（KL Divergency）约束的正则项。而本小节的优化问题直接变为：

$$\mathcal{L}^{\epsilon}(\boldsymbol{p}, \boldsymbol{q}) = \inf_{T \in \Pi[\boldsymbol{p}, \boldsymbol{q}]} < C, T > -\epsilon \mathbf{H}(T)$$

其中 $\mathbf{H}(T) \overset{\text{def.}}{=} -\sum_{i,j} T_{i,j}(\log(T_{i,j}) - 1)$.

Remark1：(强)$\lambda-$ 凸函数 $f(x)$ 是指 $f(x) - \frac{\lambda}{2}x^2$ 为（强）凸函数。由于-H(T) 的 Hessian 矩阵 $\nabla^2(-H) = diag(1/T_{i,j}) \succeq 1 (T_{i,j} \leq 1)$，因此-H(T) 为 (强)1-凸函数。上述目标函数为强 $\epsilon-$ 凸函数，可行解集也是凸集，因此上述优化问题有唯一解。这也能体现熵正则的原因：加一个凸函数使得目标函数性能（凸性）更好，利于优化。

Remark2：记 T^{ϵ} 为上述熵正则的最优传输方案，则随 $\epsilon \to 0, T^{\epsilon} \to T^*$, 其中，$T^*$ 为原始 Kantorovich 问题的最优传输方案解集中熵最大者。此外，当 $\epsilon \to \infty$ 时 $T^{\epsilon} = \boldsymbol{p}\boldsymbol{q}^{\mathsf{T}}$。

以可行解集为三角形看熵正则项对求解的影响，如图3所示。$\epsilon = 0$，也就是没有正则项，目标函数为线性函数，随着 ϵ 增加，目标函数逐渐变凸。

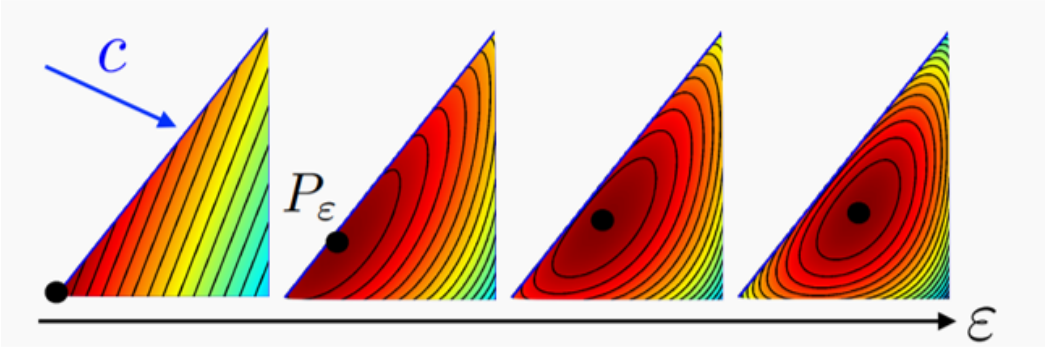


图 3: 熵正则系数 ϵ 的变化对上述优化问题求解的影响

3.2.1 熵正则与 KL 散度

任给一个凸函数 $\phi(x)$，可定义 Bregman 散度：

$$B_\phi(x, y) \overset{\text{def.}}{=} \phi(x) - \phi(y) - < \nabla\phi(y), x - y >$$

对 x 求导有：

$$\nabla_x B_\phi(x, y) = \nabla\phi(x) - \nabla\phi(y)$$

取 $\phi(x) = x\log(x)$. 可得 Bregman 散度的特例——KL 散度：

$$\mathbf{KL}(x|y) \overset{\text{def.}}{=} x\log\frac{x}{y} - x + y$$

对 x 求导有:

$$\nabla_x KL(x|y) = \log(x) - \log(y)$$

KL 散度常用于度量两个分布间的距离。在离散最优传输下，KL 散度形式如下：

$$\mathbf{KL}(P||K) \overset{\text{def.}}{=} \sum_{i,j} P_{i,j}\log\left(\frac{P_{i,j}}{K_{i,j}}\right) - P_{i,j} + K_{i,j}$$

熵正则问题的最优解 T^ϵ 为传输成本 C 的吉布斯核在 KL 散度意义下向区域 $\Pi[\boldsymbol{p}, \boldsymbol{q}]$ 的投影，即有

$$T^\epsilon = \text{Proj}^{\mathbf{KL}}_{\Pi[\boldsymbol{p},\boldsymbol{q}]}(K) \overset{\text{def.}}{=} \underset{T\in\Pi[\boldsymbol{p},\boldsymbol{q}]}{\text{argmin}}\,\mathbf{KL}(T||K), K_{i,j} = e^{-\frac{C_{i,j}}{\epsilon}}$$

证明： 记 $g(T)=\mathbf{KL}(T||K), K_{i,j}=e^{-\frac{C_{i,j}}{\epsilon}}$，

$$\begin{aligned} g(T) &= \sum_{i,j}[T_{ij}\log(\frac{T_{ij}}{K_{ij}})-T_{ij}+K_{ij}] \\ &= \sum_{i,j}T_{ij}[\log(T_{ij})+\frac{C_{ij}}{\epsilon}-1]+K_{ij} \\ &= \sum_{i,j}T_{ij}\frac{C_{ij}}{\epsilon}+T_{ij}(\log(T_{ij})-1)+K_{ij} \\ &= \frac{1}{\epsilon}[<C,T>-\epsilon\mathbf{H}(T)]+K \end{aligned}$$

因此，$\operatorname{argmin} g(T) \iff \operatorname{argmin} <C,T>-\epsilon\mathbf{H}(T)$ ■

3.2.2 熵正则问题的对偶性形式与 Sinkhorn 迭代

熵正则问题对应的 lagrange 函数如下

$$\mathcal{L}(T,\boldsymbol{f},\boldsymbol{g})=<C,T>-\epsilon\mathbf{H}(T)-<\boldsymbol{f},T\mathbb{1}_{n_2}-\boldsymbol{p}>-<\boldsymbol{g},T^{\mathsf{T}}\mathbb{1}_{n_1}-\boldsymbol{q}>$$

对 T 求导有：

$$\frac{\partial\mathcal{L}(T,\boldsymbol{f},\boldsymbol{g})}{\partial T_{i,j}}=C_{i,j}+\epsilon\log(T_{i,j})-\boldsymbol{f}_i-\boldsymbol{g}_j=0$$

由此可得

$$T_{i,j}=e^{\boldsymbol{f}_i/\epsilon}e^{-\boldsymbol{C}_{i,j}/\epsilon}e^{\mathbf{g}_j/\epsilon}$$

记 $u_i=e^{\boldsymbol{f}_i/\epsilon}, v_j=e^{\boldsymbol{g}_j/\epsilon}$，由此可得最优传输方案 $T^{\epsilon}=\operatorname{diag}(\boldsymbol{u})\mathbf{K}\operatorname{diag}(\boldsymbol{v})$，同时还需满足约束条件 (或 $\frac{\partial\mathcal{L}(T,\boldsymbol{f},\boldsymbol{g})}{\partial\boldsymbol{f}},\frac{\partial\mathcal{L}(T,\boldsymbol{f},\boldsymbol{g})}{\partial\boldsymbol{g}}=0$)：

$$\operatorname{diag}(\boldsymbol{u})\mathbf{K}\operatorname{diag}(\boldsymbol{v})\mathbb{1}_m=\boldsymbol{p},\quad \text{and}\quad \operatorname{diag}(\mathbf{v})\mathbf{K}^{\top}\operatorname{diag}(\boldsymbol{u})\mathbb{1}_n=\boldsymbol{q}$$

由于 $\operatorname{diag}(\boldsymbol{v})\mathbb{1}_m=\boldsymbol{v}$,$\operatorname{diag}(\boldsymbol{u})\mathbb{1}_n=\boldsymbol{u}$ 由此可得：

$$\boldsymbol{u}\odot(\mathbf{K}\boldsymbol{v})=\boldsymbol{p}\quad\text{and}\quad\boldsymbol{v}\odot(\mathbf{K}^{\mathrm{T}}\boldsymbol{u})=\boldsymbol{q}$$

其中 $\odot$ 为 *Hadamard product*，表示逐点相乘之意。换一下形式即得 Sinkhorn 迭代算法：

$$\boldsymbol{u}^{(\ell+1)}\overset{\text{def.}}{=}\frac{\boldsymbol{p}}{\mathbf{K}\boldsymbol{v}^{(\ell)}}\quad\text{and}\quad\boldsymbol{v}^{(\ell+1)}\overset{\text{def.}}{=}\frac{\boldsymbol{q}}{\mathbf{K}^{\mathrm{T}}\boldsymbol{u}^{(\ell+1)}}$$

Remark: Sinkhorn 算法具有并行化处理的优势。比如有 N 个最优传输问题，那么就可以分别求出每个问题的 K 矩阵，然后把他们对角式的拼接成一个更大的矩阵，同

样的把源概率密度和目标概率密度 $\boldsymbol{p}, \boldsymbol{q}$ 拼成更大的向量，就可以 N 个问题视为一个大问题，一次求解。

Remark: 可以对 $\boldsymbol{v}^0 \geq 0$ 进行任意的初始化，因为从上述的计算公式可以看出，虽然最后的 $\boldsymbol{u}, \boldsymbol{v}$ 结果不同，但是在相差一个常向量逐点相乘（另一个为相除）意义下一致。因为最终的传输方案结果一样。

3.2.3 Log 域下的 Sinkhorn 算法

为保证数值稳定性，防止因为 ϵ 过小导致的数值精度问题以防影响其收敛 ($K = e^{-\frac{C}{\epsilon}}$, 可能其中有一些项趋近 0，然而他是放到分母上来的。)，可直接对上述 Sinkhorn 迭代取 log。得

$$\mathbf{f}^{(\ell+1)} = \varepsilon \log \boldsymbol{p} - \varepsilon \log \left(\mathbf{K} e^{\mathbf{g}^{(\ell)}/\varepsilon} \right), \mathbf{g}^{(\ell+1)} = \varepsilon \log \boldsymbol{q} - \varepsilon \log \left(\mathbf{K}^{\mathrm{T}} e^{\mathbf{f}^{(\ell+1)}/\varepsilon} \right)$$

恢复的方式：

$$(\mathbf{f}^{(\ell)}, \mathbf{g}^{(\ell)}) = \varepsilon(\log(\mathbf{u}^{(\ell)}), \log(\mathbf{v}^{(\ell)}))$$

下面我们用 *soft-min* 进行改写:

$$\min_{\varepsilon} \mathbf{z} = -\varepsilon \log \sum_i e^{-\mathbf{z}_i/\varepsilon}$$

会发现 $\mathbf{f}^{(\ell+1)}, \mathbf{g}^{(\ell+1)}$ 刚好可以改写成 *soft-min* 的形式：

$$\mathbf{f}^{(\ell+1)} = \mathrm{Min}_{\varepsilon}^{\mathrm{row}}(\mathbf{C} - \mathbb{1}_n \mathbf{g}^{(\ell)^{\mathrm{T}}}) + \varepsilon \log \boldsymbol{p}$$
$$\mathbf{g}^{(\ell+1)} = \mathrm{Min}_{\varepsilon}^{\mathrm{col}}(\mathbf{C} - \mathbf{f}^{(\ell)} \mathbb{1}_m^{\mathrm{T}}) + \varepsilon \log \boldsymbol{q}$$

Note: 以 $\mathbf{f}^{(\ell+1)}$ 为例，它的第 i 个分量 $\mathbf{f}_i^{(\ell+1)}$ 中有一项为 K 的第 i 行与 g^{ℓ} 的内积 $-\log \sum_j e^{-\frac{C_{ij}-g_j^{\ell}}{\epsilon}}$ 的形式，所以 $\mathbf{f}^{(\ell+1)}, \mathbf{g}^{(\ell+1)}$ 对应可写成如上形式。具体的：$\mathrm{Min}_{\varepsilon}^{\mathrm{row}}$ 是对行 (也就是遍历每行的各列) 求 $\min_{\epsilon}$, 对应 *numpy* 的 *axis=1* 命令；$\mathrm{Min}_{\varepsilon}^{\mathrm{col}}$ 是对列 (也就是遍历每列的各行) 求 $\min_{\epsilon}$, 对应 *numpy* 的 *axis=0* 命令。

既然用上了 *soft-min*, 那必然可以用 *soft-min* 特用的数值稳定性方法：

$$\min_{\varepsilon} \mathbf{z} = \underline{z} - \varepsilon \log \sum e^{-(\mathbf{z}_i - \underline{z})/\varepsilon}, \underline{z} = \min z$$

Note:*soft-min* 与 *soft-max* 正好相对，只不过一个取正一个取负。深度学习分类问题常用的 softmax 算子 $\frac{e^{x_i}}{\sum_j e^{x_j}}$ 的分母取 log 就相当于 ϵ 取 1 下的软最大。**Note:** 如上可实现数值稳定的原因保证 $\mathbf{z} = \underline{z} > 0$, 以防止 $z < 0, e^{-z/\epsilon}$ 造成数值上溢，另一方面 $\mathbf{z} = \underline{z} > 0$ 保证了 log 里面至少有一个 $\log 1$, 防止 $\log 0$ 的下溢。类似的，softmax 则减去 $\max z$。

当然 softmax 后还有用交叉熵损失函数对每一项取 log, 还有可能有 $\log 0$ 出现，那其实就是对上一步用 $\max z$ 后，取 log 展开，会获得 $x_i - M - \log(\sum_j e^{(x_j - M)}), M = \max x$. 这样就纯稳定了，此时 log 里面至少有一个 $\log 1$, 也没其他干扰了。

因此可以用最小值改写上面的式子：

$$\mathbf{f}^{(\ell+1)} = \boldsymbol{m}_1 + \mathrm{Min}_{\varepsilon}^{\mathrm{row}}(\mathbf{C} - \mathbb{1}_n \mathbf{g}^{(\ell)^{\mathrm{T}}} - \boldsymbol{m}_1) + \varepsilon \log \boldsymbol{p}, \boldsymbol{m}_1 = \mathrm{Min}(\mathbf{C} - \mathbb{1}_n \mathbf{g}^{(\ell)^{\mathrm{T}}}, axis = 1)$$
$$\mathbf{g}^{(\ell+1)} = \boldsymbol{m}_2 + \mathrm{Min}_{\varepsilon}^{\mathrm{col}}(\mathbf{C} - \mathbf{f}^{(\ell)} \mathbb{1}_m^{\mathrm{T}} - \boldsymbol{m}_2) + \varepsilon \log \boldsymbol{q}, \boldsymbol{m}_2 = \mathrm{Min}(\mathbf{C} - \mathbb{1}_n \mathbf{g}^{(\ell)^{\mathrm{T}}}, axis = 0)$$

当然，可以不用最小值，而是用上一步的迭代结果来保证数值稳定性，这就是最终的 log 域下的 sinkhorn 算法：

$$\mathbf{f}^{(\ell+1)} = \mathrm{Min}_{\varepsilon}^{\mathrm{row}}\left(\mathbf{S}(\mathbf{f}^{(\ell)}, \mathbf{g}^{(\ell)})\right) + \mathbf{f}^{(\ell)} + \varepsilon \log(\boldsymbol{p}),$$
$$\mathbf{g}^{(\ell+1)} = \mathrm{Min}_{\varepsilon}^{\mathrm{col}}\left(\mathbf{S}(\mathbf{f}^{(\ell+1)}, \mathbf{g}^{(\ell)})\right) + \mathbf{g}^{(\ell)} + \varepsilon \log(\boldsymbol{q}),$$
$$\mathbf{S}(\mathbf{f}, \mathbf{g}) = (\mathbf{C}_{i,j} - \mathbf{f}_i - \mathbf{g}_j)_{i,j}$$

Remark: 上述 log 域的 Sinkhorn 由于需要对每行，每列处理，似乎没法直接并行。一个替代的思路就是生成两个大矩阵，一个大矩阵由每个问题的 $\mathbf{S}(\mathbf{f}, \mathbf{g})$ 矩阵延行拼接，只处理对行求和（axis=1），另一个大矩阵由每个问题的 $\mathbf{S}(\mathbf{f}, \mathbf{g})$ 矩阵延列拼接，只处理对列求和（axis=0）.

下面进一步从 Kantorvich 问题（正则化）对偶问题的坐标下降算法来进一步理解 Sinkhorn 算法。

由于上一小节推导可知 $T_{i,j} = e^{\boldsymbol{f}_i/\epsilon} e^{-\boldsymbol{C}_{i,j}/\epsilon} e^{\mathbf{g}_j/\epsilon}$, 那么对熵正则问题：

$$\mathcal{L}^{\epsilon}(\boldsymbol{p}, \boldsymbol{q}) = \inf_{T \in \Pi[\boldsymbol{p}, \boldsymbol{q}]} < C, T > -\epsilon \mathbf{H}(T)$$

用 $\boldsymbol{f}, \boldsymbol{g}$ 重新改写目标函数：

$$\langle e^{\mathbf{f}/\varepsilon}, (\mathbf{K} \odot \mathbf{C}) e^{\mathbf{g}/\varepsilon} \rangle - \varepsilon \mathbf{H}(\mathrm{diag}(e^{\mathbf{f}/\varepsilon}) \mathbf{K} \mathrm{diag}(e^{\mathbf{g}/\varepsilon}))$$

对于 $-\epsilon \mathbf{H}(T) = \epsilon \langle T, \log T - \mathbb{1}_{n \times m} \rangle$ 那一项可写作

$$\langle \mathrm{diag}(e^{\mathbf{f}/\varepsilon}) \mathbf{K}\, \mathrm{diag}(e^{\mathbf{g}/\varepsilon}), \mathbf{f} \mathbb{1}_m^{\mathrm{T}} + \mathbb{1}_n \mathbf{g}^{\mathrm{T}} - \mathbf{C} - \varepsilon \mathbb{1}_{n \times m} \rangle$$
$$= -\langle e^{\mathbf{f}/\varepsilon}, (\mathbf{K} \odot \mathbf{C}) e^{\mathbf{g}/\varepsilon} \rangle + \langle \mathbf{f}, \boldsymbol{p} \rangle + \langle \mathbf{g}, \boldsymbol{q} \rangle - \varepsilon \langle e^{\mathbf{f}/\varepsilon}, \mathbf{K} e^{\mathbf{g}/\varepsilon} \rangle$$

这样获得：

$$\mathcal{L}^{\epsilon}(\boldsymbol{p}, \boldsymbol{q}) = \max_{\mathbf{f} \in \mathbb{R}^n, \mathbf{g} \in \mathbb{R}^m} \langle \mathbf{f}, \boldsymbol{p} \rangle + \langle \mathbf{g}, \boldsymbol{q} \rangle - \varepsilon \langle e^{\mathbf{f}/\varepsilon}, \mathbf{K} e^{\mathbf{g}/\varepsilon} \rangle$$

即把原问题的熵正则问题变为对偶问题的正则问题了！

记目标函数为 Q, 对上述问题分别对 $\mathbf{f},\mathbf{g}$ 求导，可得

$$\nabla|_{\mathbf{f}}Q(\mathbf{f},\mathbf{g})=\boldsymbol{p}-e^{\mathbf{f}/\varepsilon}\odot\left(\mathbf{K}e^{\mathbf{g}/\varepsilon}\right),\nabla|_{\mathbf{g}}Q(\mathbf{f},\mathbf{g})=\boldsymbol{q}-e^{\mathbf{g}/\varepsilon}\odot\left(\mathbf{K}^{\mathrm{T}}e^{\mathbf{f}/\varepsilon}\right).$$

因此可以看出（log 域）Sinkhorn 算法：

$$\mathbf{f}^{(\ell+1)}=\varepsilon\log\boldsymbol{p}-\varepsilon\log\left(\mathbf{K}e^{\mathbf{g}^{(\ell)}/\varepsilon}\right),\mathbf{g}^{(\ell+1)}=\varepsilon\log\boldsymbol{q}-\varepsilon\log\left(\mathbf{K}^{\mathrm{T}}e^{\mathbf{f}^{(\ell+1)}/\varepsilon}\right)$$

正是求解正则对偶问题的轮换坐标下降!

3.2.4 Stochastic Dual Entropy Method in Deep learning

要在 Deep learning 的背景下求解最优传输 coupling（连续形式）需考虑两个技术问题，一是过程可微分，二是选择一个合理的采用神经网络的表示方式。

很自然的方法（前面讲过了 Kantorovich 对偶问题以及前文数值解法做了铺垫），就是借助 dual 问题，把 Kantorovich 势能函数建模为神经网络，用随机梯度下降或者其他 deeplearning 中常用的优化方法优化这两个 Kantorovich 势能，最后恢复出最优传输方案。

考虑熵正则的最优传输问题：

$$\inf_{\gamma\in\Pi(\mu,\nu)}\int_{X\times Y}c(x,y)d\gamma(x,y)-\varepsilon\mathbf{H}(\gamma)$$

完全类比离散形式，或者直接用 Lagrange dual 重新算我们知道最优的传输方案具有形式：

$$\gamma_{\varepsilon}^{*}=e^{\frac{\boldsymbol{f}(\boldsymbol{x})}{\varepsilon}}e^{-\frac{c(x,y)}{\varepsilon}}e^{\frac{\boldsymbol{g}(\boldsymbol{y})}{\varepsilon}}$$

然后和上一小节一样，把这个结果带回到原问题，并用 $\boldsymbol{f}(\boldsymbol{x}),\boldsymbol{g}(\boldsymbol{y})$ 表示，形式如下:

$$\sup_{f,g}\int_{\mathcal{X}\times\mathcal{Y}}f(x)+g(y)-\varepsilon e^{\frac{f(x)+g(y)-c(x,y)}{\varepsilon}}d\mu(x)d\nu(y)$$

当然，在 deep learning 背景下，$f=f_{\theta_1}(x),g=g_{\theta_2}$ 用两个神经网络表示即可；在算具体的损失时，直接采样一堆样本算均值即可：

$$\sup_{f,g}\mathbb{E}_{(X,Y)\sim\mu\times\nu}\left[f(X)+g(Y)-\varepsilon e^{\frac{f(X)+g(Y)-c(X,Y)}{\varepsilon}}\right]$$

由于这里是 sup, 在写程序的时候注意要么在 loss 那边, 要么在梯度那边加个负号。

当然，你也可以采用半对偶的形式，有：

$$W_{\varepsilon}(\mu,\nu)=\max_{g}H_{\varepsilon}(g)=\max_{g}\int_{X}g^{c,\varepsilon}(x)d\mu+\int_{Y}g(y)d\nu-\varepsilon=\max_{g}\mathbb{E}_{X}\left[h(x,g)\right]$$

其中 $h(x,g)=g^{c,\varepsilon}(x)+\int_Y g(y)d\nu-\varepsilon,\forall\varepsilon>0, g^{c,\varepsilon}(x):=-\varepsilon\log\left[\int_Y e^{\frac{g(y)-c(x,y)}{\varepsilon}}d\nu(y)\right]$，这里由于采用熵正则，我们不必需要 g 为 c-concave 函数的要求了，直接用一个神经网络表示就行了。求出最优的 g 后，f 就相当于 $g^{c,\varepsilon}$，然后套用之前求 γ_ε^* 就行了。虽然优化过程中，积分都是期望，期望用平均来算，没啥问题，不过最后获得 $f=g^{c,\varepsilon}$ 里面有积分项，用均值来算大抵不是很优雅，所以还是用 dual 形式更好，不过我们将在后面的章节着重讨论不考虑熵正则情况下，基于 semi-dual 形式的计算，参见 Sec.3.8.

3.3 近端点法 (Proximal Point Method)

强化学习流形的 TRPO、PPO 也是类似的策略。不像上一小节的熵正则问题，近端点法本质上求解的是原始 OT 问题，但其每一步迭代又可以化为一个熵正则的 OT 问题，这里的熵为相对熵也叫 KL 散度。

近端算子的定义如下：

$$\text{prox}_f(\mathbf{x})=\text{argmin}_{\mathbf{u}\in\mathbb{E}}\left\{f(\mathbf{u})+\frac{1}{2}\|\mathbf{u}-\mathbf{x}\|^2\right\}\forall\mathbf{x}\in\mathbb{E}.$$

Remark: 可以把近端点法看作正常梯度下降的隐式框架，也就是 $\boldsymbol{x}_{k+1}=\boldsymbol{x}_k-\alpha_k\nabla f(\boldsymbol{x}_{k+1})$。因为这正是近端点法问题 $\boldsymbol{x}_{k+1}=\arg\min_{\boldsymbol{x}\in\mathbb{R}^n}\left(f(\boldsymbol{x})+\frac{1}{2\alpha_k}\|\boldsymbol{x}-\boldsymbol{x}_k\|_2^2\right)$ 的最优解。换言之，可把普通的梯度下降视为前向欧拉方法而近端梯度下降为后向欧拉方法。

将其中的欧式距离（的一半）换成更广义的 Bregman 散度 B_ϕ，可得更一般意义下的近端梯度算子：

$$\text{prox}_f^{B_\phi}(\mathbf{x})=\text{argmin}_{\mathbf{u}\in\mathbb{E}}\left\{f(\mathbf{u})+B_\phi(u,x)\right\}\forall\mathbf{x}\in\mathbb{E}.$$

利用近端点法，取 $\phi=x\log x$, 求解原始 Kantorovich 问题，则其迭代如下：

$$T^{(\ell+1)}\overset{\text{def.}}{=}\text{Prox}_{\frac{1}{\varepsilon}f}^{\mathbf{KL}}(T^{(\ell)})\overset{\text{def.}}{=}\underset{T\in\Pi[\boldsymbol{p},\boldsymbol{q}]}{\text{argmin}}\mathbf{KL}(T|T^{(\ell)})+\frac{1}{\epsilon}<C,T>$$

上述近端点迭代算法等价于求解成本函数为 $\bar{C}=C-\epsilon\log(T^\ell)$ 的熵正则 Kantorovich 问题。

证明 1： 记 $g(T)=\epsilon\mathbf{KL}(T|T^{(\ell)})+<C,T>$, 则 $T^{(\ell+1)}=\underset{T\in\Pi[\boldsymbol{p},\boldsymbol{q}]}{\text{argmin}}g(T)$.

$$\begin{aligned}
g(T)&=\epsilon\mathbf{KL}(T|T^{(\ell)})+<C,T>\\
&=<C,T>+\epsilon(\sum_{i,j}T_{i,j}\log(\frac{T_{i,j}}{T_{i,j}^\ell})-T_{i,j}+T_{i,j}^\ell)\\
&=<C-\epsilon\log(T^\ell),T>+\epsilon<\log(T)-1,T>+\epsilon T^\ell\\
&=<C-\epsilon\log(T^\ell),T>-\epsilon\mathbf{H}(T)+\epsilon T^\ell
\end{aligned}$$

因此，$\operatorname{argmin} g(T) \iff \operatorname{argmin} < C - \epsilon \log(T^\ell), T > -\epsilon \mathbf{H}(T)$ ■

此时 $K = e^{-\frac{\bar{C}}{\epsilon}} = T^\ell \odot e^{-\frac{C}{\epsilon}}$. 且有：

$$
\begin{aligned}
T^{(\ell+1)} &= \operatorname{diag}(\mathbf{u}^{(\ell)})(e^{-\frac{\mathbf{C}}{\varepsilon}} \odot T^{(\ell)})\operatorname{diag}(\mathbf{v}^{(\ell)}) \\
&= \operatorname{diag}(\mathbf{u}^{(\ell)} \odot \cdots \odot \mathbf{u}^{(0)}) e^{-\frac{(\ell+1)\mathbf{C}}{\varepsilon}} \odot T^{(0)}) \operatorname{diag}(\mathbf{v}^{(\ell)} \odot \cdots \odot \mathbf{v}^{(0)})
\end{aligned}
$$

可取初始化 $T^{(0)} = \boldsymbol{p}\boldsymbol{q}^{\mathbf{T}}$。若取初始化为 $T^{(0)} = \mathbb{1}_{n_1}\mathbb{1}_{n_2}^\top$, 则 $T^{(\ell+1)} = \operatorname{diag}(\mathbf{u}^{(\ell)} \odot \cdots \odot \mathbf{u}^{(0)}) e^{-\frac{(\ell+1)\mathbf{C}}{\varepsilon}}) \operatorname{diag}(\mathbf{v}^{(\ell)} \odot \cdots \odot \mathbf{v}^{(0)})$, 则 $K = e^{-\frac{C}{\epsilon/(\ell+1)}}$, 即等价于求解一熵正则系数为 $\frac{\epsilon}{\ell+1}$ 的熵正则问题!当然,你用第一种初始化结果也是一样的,毕竟 $\boldsymbol{p}\boldsymbol{q}^{\mathbf{T}} = \operatorname{diag}(\boldsymbol{p})\mathbb{1}_{n_1}\mathbb{1}_{n_2}^\top \operatorname{diag}(\boldsymbol{q})$, 只不过把两侧的 diag 吸收到待求的 $\boldsymbol{u}, \boldsymbol{v}$ 中，不影响优化。

特别的，如果采用变正则系数的优化, 形如：

$$
Q^{k+1} \triangleq \underset{Q \in \Pi_{\boldsymbol{p},\boldsymbol{q}}}{\operatorname{argmin}} \langle C, Q \rangle + \frac{1}{\gamma_k} \mathrm{KL}(Q, Q_k)
$$

初始化为 $Q^0 = \boldsymbol{p}\boldsymbol{q}^\top$ 或 $\mathbb{1}_{n_1}\mathbb{1}_{n_2}^\top$, 那么根据上面的推导，则迭代过程等价于优化一系列系数为 $\{\varepsilon_k\}$ 的熵正则问题：

$$
Q^{k+1} \triangleq \underset{Q \in \Pi_{\boldsymbol{p},\boldsymbol{p}}}{\operatorname{argmin}} \langle C, Q \rangle - \varepsilon_k H(Q)
$$

其中 $\varepsilon_k \triangleq (\sum_{j=0}^{k} \gamma_j)^{-1}$.

证明 2（基于镜像梯度下降）： 根据 Sec.4.2.1, 更新上述最优问题的更新步为：

$$
\begin{aligned}
T^{(\ell+1)} &= \operatorname{Proj}^{\mathbf{KL}}_{\Pi[\boldsymbol{p},\boldsymbol{q}]}(T^{(\ell)} \odot e^{-\tau \nabla g(T^{(\ell)})}) \\
&= \operatorname{Proj}^{\mathbf{KL}}_{\Pi[\boldsymbol{p},\boldsymbol{q}]}(T^{(\ell)} \odot e^{-\tau C})
\end{aligned}
$$

令 $K = T^{(\ell)} \odot e^{-\tau C} = exp(-\frac{\bar{C}}{\epsilon})$, 可得 $\bar{C} = -\epsilon \log(K) = \epsilon\tau C - \epsilon \log(T^{(\ell)})$。即采用 Mirror Descent Algorithm, 近端点算法的每次迭代等价于成本函数为 $\bar{C} = \epsilon\tau C - \epsilon \log(T^{(\ell)})$, 熵正则系数为 ϵ 的熵正则最优传输问题。若取 $\epsilon\tau = 1$, 则 $\bar{C} = C - \epsilon \log(T^{(\ell)})$ ■

证明 3（Larangian dual）：

近端点迭代 (第 $\ell+1$ 步) 问题对应的 lagrange 函数如下

$$
\mathcal{L}(T, \boldsymbol{f}, \boldsymbol{g}) = < C, T > + \epsilon \mathbf{KL}(T||T^\ell) - < \boldsymbol{f}, T\mathbb{1}_{n_2} - \boldsymbol{p} > - < \boldsymbol{g}, T^{\mathsf{T}}\mathbb{1}_{n_1} - \boldsymbol{q} >
$$

对 T 求导有：

$$
\frac{\partial \mathcal{L}(T, \boldsymbol{f}, \boldsymbol{g})}{\partial T_{i,j}} = C_{i,j} + \epsilon \log(T_{i,j}) - \epsilon \log(T^\ell_{i,j}) - \boldsymbol{f}_i - \boldsymbol{g}_j = 0
$$

由此可得

$$
T_{i,j} = e^{\boldsymbol{f}_i/\epsilon} T^\ell_{i,j} e^{-\boldsymbol{C}_{i,j}/\epsilon} e^{\mathbf{g}_j/\epsilon}
$$

记 $u_i = e^{\boldsymbol{f}_i/\epsilon}, v_j = e^{\boldsymbol{g}_j/\epsilon}$, 由此可得最优传输方案, 也就是下一步的 $T^{\ell+1} = \mathrm{diag}(\boldsymbol{u})T^{\ell}\odot \mathbf{K}\,\mathrm{diag}(\boldsymbol{v})$, 同时还需满足约束条件 $(T^{\ell+1}\mathbb{1}_{n_2} - \boldsymbol{p} = 0, {T^{\ell+1}}^{\top}\mathbb{1}_{n_1} - \boldsymbol{q} = 0)$, 这就引出对 $\boldsymbol{u}$ $\boldsymbol{v}$ 的更新。从上述形式上可以看出，这其实对应成本函数 $\bar{C} = C - \epsilon\log(T^{\ell})$ 的 Entropic OT 问题。 ■

3.4 Bregman 迭代算法

Bregman 迭代将可行解集 $\Pi[\boldsymbol{p}, \boldsymbol{q}]$ 松弛为两个:

$$\Pi_{\mathbf{p}}^1 \overset{\text{def.}}{=} \{T : T\mathbb{1}_m = \boldsymbol{p}\} \quad \text{and} \quad \Pi_{\mathbf{q}}^2 \overset{\text{def.}}{=} \left\{T : T^{\mathrm{T}}\mathbb{1}_n = \boldsymbol{q}\right\}$$

有 $\Pi[\boldsymbol{p}, \boldsymbol{q}] = \Pi_{\mathbf{p}}^1 \cap \Pi_{\mathbf{q}}^2$, 将熵正则最优传输问题 $\mathrm{Proj}^{\mathbf{KL}}_{\Pi[\boldsymbol{p},\boldsymbol{q}]}(K)$ 转为在,$\Pi_{\mathbf{p}}^1, \Pi_{\mathbf{q}}^2$ 这两个更大空间上交替迭代:

$$T^{(\ell+1)} \overset{\text{def.}}{=} \mathrm{Proj}^{\mathbf{KL}}_{\Pi_{\mathbf{p}}^1}(T^{(\ell)}) \quad \text{and} \quad T^{(\ell+2)} \overset{\text{def.}}{=} \mathrm{Proj}^{\mathbf{KL}}_{\Pi_{\mathbf{q}}^2}(T^{(\ell+1)})$$

其中 $T^{(0)} = K = e^{-\frac{C}{\epsilon}}$.

根据文献知:$\forall n > 0, \quad \gamma^{(n)} \overset{\text{def.}}{=} \mathrm{Proj}^{\mathbf{KL}}_{\Pi^n}(\gamma^{(n-1)})$, 可以证明,$\gamma^{(n)} \to \mathrm{Proj}^{\mathrm{KL}}_{\Pi[\boldsymbol{p},\boldsymbol{q}]}(K) \quad \text{as } n \to \infty$, 收敛到唯一解。具体的解法，也就是 KL 投影运算会根据约束条件有相应的解析形式。这里有:

$$\mathrm{Proj}^{\mathrm{KL}}_{\Pi_{\mathbf{p}}^1}(\bar{\gamma}) = \mathrm{diag}\left(\frac{\boldsymbol{p}}{\bar{\gamma}\mathbb{1}_m}\right)\bar{\gamma} \quad \textit{and} \quad \mathrm{Proj}^{\mathrm{KL}}_{\Pi_{\mathbf{q}}^2}(\bar{\gamma}) = \bar{\gamma}\mathrm{diag}\left(\frac{\boldsymbol{q}}{\bar{\gamma}^T\mathbb{1}_n}\right)$$

你会发现，上述迭代公式本质上就是按照边缘分布（给定的源域目标域的分布）在不断通过缩放的形式更正传输方案，直到求解的结果满足约束条件 (也就是两个边缘分布的约束，此时更新的系数为单位阵了)。

这里具体的推导过程为：构建拉格朗日函数求出最优解，然后最优利用约束条件进行正则化以使解可行。以第一个为例描述其推导过程。

证明： 待求问题为:

$$T^{(\ell+1)} = \arg\min_{T\in\Pi_{\mathbf{p}}^1} \mathbf{KL}(T|T^{(\ell)})$$

, 其拉格朗日函数为: $\mathcal{L}(T, \lambda, \omega) = B_{\phi}(T, T^{(\ell)}) - \lambda^{\top}(T\mathbb{1}_m - \boldsymbol{p}) - \omega \odot T$. 令

$$\nabla_T\mathcal{L} = \nabla\phi(T) - \nabla\phi(T^{\ell}) - \omega - \lambda\mathbb{1}_m^{\top} = 0$$

由于 $\phi = x\log(x)$, 因此有:

$$\log(T) = \log(T^{\ell}) + \omega + \lambda\mathbb{1}_m^{\top}$$

, 因此, $T = A \odot T^{\ell}$ 的形式。由于 $T\mathbb{1}_m = \boldsymbol{p}$, 因而 $A \odot T^{\ell}\mathbb{1}_m = \boldsymbol{p}$, 可得 $A = \mathrm{diag}(\boldsymbol{p} \oslash (T^{\ell}\mathbb{1}_m))$, 因此最优解为 $T = \mathrm{diag}(\frac{\boldsymbol{p}}{T^{\ell}\mathbb{1}_m})T^{(\ell)}$ ■

Remark: IBP(迭代 Bregman 投影) 算法收敛是因为两个投影空间是仿射凸集。一般地，当可行解集可分解为多个闭凸集（用于迭代执行投影操作）之交时，IBP 算法不一定会收敛到目标空间，下一小节我们介绍他的一个拓展版本。

3.5 Dykstra' s Algorithm

本节介绍 Dykstra' s Algorithm，作为上一小节 IBP 算法的扩展，可以保证当凸的可行解集可以分解为多个闭凸集之交时，通过每次在闭凸集做投影，最终可收敛到可行解集（各凸集之交）。当然本小节的方法亦应用在 Sec.3.6。

3.6 Low Rank Shinhorn

3.7 Fast Sinkhorn for Wassertein-1

3.7.1 Fast Sinkhorn I

3.7.2 Fast Sinkhorn II

3.8 Neural Optimal Transport Solvers for Transport Map

参考Optimal transport mapping via input convex neural networks极其后续工作，如Brandon Amos处理 legendre dual 的On amortizing convex conjugates for optimal transport;ICNN 神经网络在求解 Wasserstein Barycenter(见 Sec.7.3) 以及 Wasserstein Gradient Flow(见 Sec.9.3.3) 亦有应用

对于考虑一般成本函数，有的Marco Cuturi等人的The Monge Gap 以及Alexander Korotin的Neural Optimal Transport和Neural Optimal Transport with General Cost Functionals;

最后介绍最近Yongxin Chen团队的Displacement Interpolation Optimal Transport Model

3.8.1 欧式距离下参数化 Kantorovich 势能函数

我们已经在 Sec.3.2.4一窥 Neural Optimal Transport Solvers. 本节主要聚焦于学成本函数为欧氏距离下的最优传输映射。显然原理已经在第二章 Sec.2.3.3都给出了:

所求 Monge 问题

$$\inf\left\{M(T) := \frac{1}{2}\int_X |x - T(x)|^2 d\mu, \quad T_{\#}\mu = \nu\right\}$$

通过 Kantorovich semi dual 转化为：

$$\inf\left\{\int_X \varphi d\mu + \int_Y \varphi^* d\nu : \tilde{\varphi}(x)\ is\ convex\ function\rangle\right\}$$

其中 $\varphi^*(y) = \sup_{x\in X} < x, y > -\varphi(x)$ 为 Legendre transform. 最优传输映射 $T^\star = \nabla\varphi^\star(x)$.

本节主要是通过学 $\varphi^\star$ 来得到 $T^\star$。由于并不考虑 Entropic Regulazation, 所以主要的技术难点是用神经网络表示凸函数，以及用神经网络表示或近似凸函数的 Legendre 对偶。主要参考OTT-jax工具箱用的方法

3.8.2 任意成本函数下直接参数化传输映射

本节主要介绍The Monge Gap: A Regularizer to Learn All Transport Maps这篇文章

3.9 流体力学视角下的 Benamou-Brenier 算法

对于 Sec.2.6.2最终得到的 W_2 形式, 本节介绍用于优化此问题的 Benamou-Brenier 算法。

该方法基于如下观察：

$$\frac{\|J\|_2^2}{2\rho} = \begin{cases} \sup_{a\in\mathbb{R}, b\in\mathbb{R}^n} & a\rho + b^\top J \\ \text{s.t.} & a + \frac{\|b\|_2^2}{2} \le 0 \end{cases}$$

这是因为：利用拉格朗日乘子法，$L(a,b,f) = a\rho + b^\top J + f*(a + \frac{\|b\|_2^2}{2}), \frac{\partial L}{\partial a} = \rho + f = 0, \frac{\partial L}{\partial b} = J + bf = 0, \frac{\partial L}{\partial f} = a + \frac{\|b\|_2^2}{2} = 0.$

带入上述关系，目标函数变为：

$$a\rho + b^\top J = a\rho - b^\top bf = (a + \|b\|_2^2)\rho = \frac{\|b\|_2^2}{2}\rho = \frac{(b\rho)^\top(b\rho)}{2\rho} = \frac{\|J\|_2^2}{2\rho}$$

借助上述关系，原问题等价于求解下属优化问题：

$$\begin{aligned}\inf_{J,\rho}\sup_{a,b}&\int_0^1\int_{\mathbb{R}^n}[a(x,t)\rho(x,t)+b(x,t)^\top J(x,t)]dxdt\\ \text{s.t.}\quad&\rho(x,0)\equiv\rho_0(x)\forall x\in\mathbb{R}^n\\&\rho(x,1)\equiv\rho_1(x)\forall x\in\mathbb{R}^n\\&\frac{\partial\rho(x,t)}{\partial t}=-\nabla\cdot J(x,t)\forall x\in\mathbb{R}^n,t\in(0,1)\\&a(x,t)+\frac{\|b(x,t)\|_2^2}{2}\le 0\forall x\in\mathbb{R}^n,t\in(0,1).\end{aligned}$$

因此原问题等价于下属问题：

$$\begin{aligned}\inf_{J,\rho}\sup_{a,b,\phi}&\int_0^1\int_{\mathbf{R}^n}\left[a(x,t)\rho(x,t)+b(x,t)^\top J(x,t)+\phi(x,t)\left(\frac{\partial\rho(x,t)}{\partial t}+\nabla\cdot J(x,t)\right)\right]dxdt\\&+\int_{\mathbf{R}^n}[\phi(x,1)(\rho_1(x)-\rho(x,1))-\phi(x,0)(\rho_0(x)-\rho(x,0))]dx\\\text{s.t.}\quad&a(x,t)+\frac{\|b(x,t)\|_2^2}{2}\le 0,\forall x\in\mathbb{R}^n,t\in(0,1).\end{aligned}$$

对连续方程对应的那两项用分布积分可得

$$\int_0^1\phi(x,t)\frac{\partial\rho(x,t)}{\partial t}dt=[\rho(x,1)\phi(x,1)-\rho(x,0)\phi(x,0)]-\int_0^1\rho(x,t)\frac{\partial\phi(x,t)}{\partial t}dt$$

加上边界上动量为 0,

$$\int_0^1\phi(x,t)\nabla\cdot J(x,t)dt=0-\int_0^1 J(x,t)\nabla\phi(x,t)\partial tdt$$

因此原问题目标函数化为：

$$\begin{aligned}\int_{\mathbf{R}^n}\Bigg\{\int_0^1(\rho(x,t)[\,a(x,t)-\frac{\partial\phi(x,t)}{\partial t}\Big]+J(x,t)^\top[b(x,t)-\nabla\phi(x,t)]dt\\-\phi(x,0)\rho_0(x)+\phi(x,1)\rho_1(x))dx\end{aligned}$$

另一种等价写法为：记 $m=(\rho,J),\xi=(a,b),\nabla_{t,x}\phi=(\partial_t\phi,\nabla_x\phi)$

$$G(\phi):=\int_\Omega\phi\,(1,x)\,\mathrm{d}\nu-\int_\Omega\phi\,(0,x)\,\mathrm{d}\mu=\int_{\Omega(\mathbb{R}^n)}-\phi(x,0)\rho_0(x)+\phi(x,1)\rho_1(x))dx$$

$$\langle m,\xi\rangle:=\int_{\mathbb{R}^n}\int_0^1(a(x,t)\rho(x,t)+b(x,t)^\top J(x,t))dtdx$$

则优化问题可写作：

$$\min_m\sup_{\xi,\phi:\xi\in K_\xi}\langle\xi-\nabla_{t,x}\phi,m\rangle+G(\phi)$$

对于约束项，令：

$$F(\xi) := \begin{cases} 0 & a(x,t) + \frac{\|b(x,t)\|_2^2}{2} \leq 0 \forall x \in \mathbb{R}^n, t \in (0,1) \\ +\infty & \text{otherwise.} \end{cases}$$

则上述原问题可改写为如下形式 ($\sup(-f) = -\inf f$)：

$$-\sup_{m} \inf_{\xi,\phi} \langle \nabla_{t,x}\phi - \xi, m\rangle - G(\phi) + F(\xi)$$

尽管上式为原问题，但是我们可以把他视为 Dual 问题（省去那个负号），m 就可以看作 lagrange 乘子（做用在约束 $\nabla_{t,x}\phi = \xi$），原变量为 ξ, ϕ.

这里采用 *augmented Lagrangian* 法（对 $\sup_{\xi,\phi}$, 引入了 $-\frac{\tau}{2}|\xi - \nabla_{t,x}\phi|^2$ 增加了函数凹性，利于求解最大值，但不改变鞍点。正如对 min 问题加入正则项 $+\frac{\tau}{2}|x - x^k|^2$, 可增加凸性)：

$$\min_{m} \sup_{\xi,\phi:\xi\in K_\xi} \langle \xi - \nabla_{t,x}\phi, m\rangle + G(\phi) - \frac{\tau}{2}|\xi - \nabla_{t,x}\phi|^2$$

记 $L_r(\phi, \xi, m) := F(\xi) - G(\phi) + \langle m, \nabla_{x,t}\phi - \xi\rangle + \frac{\tau}{2}\langle \nabla_{x,t}\phi - \xi, \nabla_{x,t}\phi - \xi\rangle$

因此，根据上述最优问题的表达式，迭代过程就是取 ξ, ϕ 以最小化 L_r 后取 m 最大化 L_r。最终的迭代框架为:

$$\phi^{\ell+1} \leftarrow \arg\min_{\phi} L_r(\phi, \xi^\ell, m^\ell)$$

$$\xi^{\ell+1} \leftarrow \arg\min_{\xi\in K_\xi} L_r(\phi^{\ell+1}, \xi, m^\ell)$$

$$m^{\ell+1} \leftarrow m^\ell - \tau(\xi^{\ell+1} - \nabla_{x,t}\phi^{\ell+1})$$

前两步（对 ϕ, ξ）写开为：

$$\phi^{\ell+1} \leftarrow \tau\Delta_{t,x}\phi = \nabla\cdot(\tau\xi_k - m_k)$$

$$\xi^{\ell+1} \leftarrow \text{Proj}_{K_\xi}[\nabla_{t,x}\hat{\phi}_{k+1} + \frac{1}{\tau}m_k]$$

其中第二项就是对 ξ 求导的一阶条件，并向约束集投影。第一步转化为一个求解 Poission 方程的问题。这一步主要应用了 *Dirichlet* 积分：即求解如下泛函极值问题（固定边界)：

$$\min_{u} \int\int \frac{1}{2}\|\nabla u\|^2 + f\cdot u dx$$

等价于求解 Poision 方程：$\Delta u = f$.

对 ϕ, 其更新依靠求解：

$$\max_{\phi} \quad -\langle \nabla_{t,x}\phi, m_k\rangle + G(\phi) - \frac{\tau}{2}||\xi_k - \nabla_{t,x}\phi||^2$$

把第一项利用分布积分（J 在边界处为 0, 对 ρ 正好和 $G(\phi)$ 抵消）化为

$$\max_{\phi} \quad \langle \phi, \nabla m_k \rangle - \frac{\tau}{2} ||\nabla_{t,x}\phi - \xi_k||^2$$

也就是

$$\min_{\phi} \quad \langle \phi, -\nabla m_k \rangle + \frac{\tau}{2} ||\nabla_{t,x}\phi - \xi_k||^2$$

把最后一项展开，用分部积分把 $- < \nabla_{t,x}\phi, \xi >$ 换成 $< \phi, \nabla\xi >$，ϕ 边界固定，就可获得如上泊松方程。

3.10 流体力学视角下的 Angenent-Hacker-Tannenbaum 算法

这个算法的主要思想就是，从一个可行解出发，通过对可行解进行调整，使得传输代价不断减小。初始传输映射 T 可以通过约束解一个泊松方程获得，具体的调整是通过流场。上一小节了解到，一个传输方案对应一个流，这个流受速度场的影响，因而把初始可行解作为流，通过增加一个向量场进而引进一个流，用新引进的流来调整初始的流，使得传输代价最小，当然新引进的流需保测度，不可影响到目标测度，即用 $T \circ (Y_t)^{-1}$ 去代替初始的 T. 其中 Y_t 就是给定某个速度场后上述一阶 ODE 的解，由于时间是可逆的（从那个 ODE 角度或 Y_t 表示从初始位置流动时间 t 后的位置角度），有 $Y_t^{-1} = Y_{-t}$. 也颇有变分思想：咋一个解上加一族满足约束的扰动，然后求导等于 0 获得最优的必要条件。

现在我们要探究用于调整传输映射的速度场 v_t 的性质。首先，由 Y_t 保测度条件得，密度场不随时间变化，即 $\rho_t \equiv f$, 根据连续方程知 $\nabla \cdot (\rho_t v_t) = 0$, 令 $w_t = \rho_t v_t$, 有 $\nabla w = 0$. 其次，我们想要通过 $T \circ Y_t^{-1}$ 来使得传输代价变小，那得研究传输代价相对于 $T \circ Y_t^{-1}$ 的导数。Monge cost 为下式：

$$M(T) = \int c(x, T(x)) d\mu_0 = \int c(x, T(x)) f(x) dx$$

设置 $T_t(x) := T \circ Y_{-t}(x)$, 令 $x' = Y_{-t}(x)$, 则 $x = Y_t(x')$, 可得 $c(x, T(Y_{-t}(x))) = c(Y_t(x'), x')$, 且上文提 Y_t 保测度，有 $f(x)dx = f(x')dx'$ ，因而

$$\begin{aligned} M(T_t) &= \int c(x, T(Y_{-t}(x))) f(x) dx = \int c(Y_t(x'), T(x')) f(x') dx' \\ &= \int c(Y_t(x), T(x)) f(x) dx \end{aligned}$$

可得

$$\begin{aligned} \frac{d}{dt} M(T_t) &= \frac{d}{dt} \int c(Y_t(x), T(x)) f(x) dx = \int \frac{d}{dt} [c(Y_t(x), T(x)) f(x) dx] \\ &= \int \nabla_x c(Y_t(x), T(x)) Y_t'(x) f(x) dx = \int \nabla_x c(Y_t(x), T(x)) v_t(x) f(x) dx \end{aligned}$$

令 $t=0$, 得到

$$\frac{d}{dt}M(T_t)|_{T_t=T}=\int\nabla_x c(x,T(x))v(x)f(x)dx=\int\nabla_x c(x,T(x))w(x)dx$$

记 $\nabla_x c(x,T(x))=\xi(x)$，则 $\frac{d}{dt}M(T_t)|_{T_t=T}=\int\xi(x)w(x)dx$. 我们的目标是通过 T_t 改进初始传输映射 T, 使得传输代价最小，因此希望 $\frac{d}{dt}M(T_t)|_{T_t=T}\leq 0$. 因此，我们为调整传输映射所加的无散流量场 w, 最优选择是 $-\xi$ 向无散场的投影, 即 $w=P[-\xi]$. 这样选择时，

$$\begin{aligned}\frac{d}{dt}M(T_t)&=\int\nabla_x c(Y_t(x),T(x))v_t(x)f(x)dx=\int\xi_t(x)w_t(x)dx\\&=-\int\xi_t(x)P[\xi_t(x)]dx\leq 0\end{aligned}$$

这说明，当流量场 w 如上选择时，流 $M(T)$ 使得 $M(T_t)$ 递减。当成本最小时，有 $P[\xi_t(x)]=0$, 即 $\xi_t(x)$ 没有无散分量，为一个梯度场（一个函数的梯度。）

Note: 由 Hodge-Helmholtz 分解定理（类似于 Brenier 极分解）知，向量场可以分解为一个梯度场和一个无散场即 $\xi=\nabla u+w,\nabla\cdot w=0$.

Remark: 我们知道当 $c(x,y)=\frac{1}{2}|x-y|^2$ 时，有 $x-T^*(x)=\nabla\phi(x)$, 即最优传输是个 T^* 是一个凸函数 $u(x)=\frac{1}{2}x^2-\phi(x)$ 的梯度。现在这个算法表示, 所能得到的最优传输映射 T^* 下，$\xi_t(x)=\nabla_x[c(x,T^*(x))]=T^*(x)=\nabla g(x)$, 但是 $g(x)$ 并不保证凸函数！不过在二维情况下，这个算法是成立的，会收敛到最优。但是更高维度仍是个开放问题!

现在我们进一步讨论 w 与 T_t.

对于 w,$w=-P[\xi]$. 为方便期间直接令 $\xi=-\nabla_x c(x,T)$, 也就是对之前定义的 ξ 直接加负号。现在由分解定理，$\xi=\nabla u+w$, 则 $w=\xi-\nabla u$，$\nabla\xi=\Delta u+\nabla w=\Delta u$, 因此 $u=\Delta^{-1}(\nabla\xi)$, 因而

$$w=P[\xi]=\xi-\nabla(\Delta^{-1}(\nabla\cdot\xi))$$

对于 T_t, 有 $T_t(Y_t(x))=T(x)$, 对 t 求偏导有

$$\frac{d}{dt}T_t(Y_t(x))=\partial_t T_t+v_t\cdot\nabla T_t=\frac{d}{dt}T(x)=0$$

即 T_t 满足 *(线性) 传输方程*:

$$\partial_t T_t+v_t\cdot\nabla T_t=0$$

根据传输方程, 在时间维度上不断更新 T_t，就是 AHT 算法。

另外，根据传输方程，以及流量场的 Neumann 边界条件也能推导出 $\frac{d}{dt}M(T_t)$.

$$\begin{aligned}\frac{d}{dt}M(T_t) &= \frac{d}{dt}\int c(x,T_t(x)))f(x)dx = \int \frac{d}{dt}[c(x,T_t(x))f(x)dx] \\ &= \int \nabla_y c(x,T_t(x))\cdot \partial_t T_t \cdot f(x)dx \\ &= -\int \nabla_y c(x,T_t(x))\cdot (\nabla T_t v_t) f(x)dx \\ &= -\int \nabla_y c(x,T_t(x))\cdot \nabla T_t \cdot w_t dx\end{aligned}$$

考察 $\nabla c(x,T_t)$, 有

$$\nabla c = \nabla_x c(x,T_t) + \nabla_y c(x,T_t)\cdot \nabla T_t$$

以及

$$\int \nabla c \cdot w = \int \nabla\cdot(cw) \overset{Stokes}{=} \int_{\partial_\Omega} cw\vec{n} \overset{Neumann}{=} 0$$

因此有 $\nabla_x c(x,T_t) = -\nabla_y c(x,T_t)\cdot \nabla T_t$, 带入上方 $\frac{d}{dt}M(T_t)$ 表达式有:

$$\frac{d}{dt}M(T_t) = \int \nabla_x c(x,T_t)\cdot w_t dx = \int \xi_t \cdot w_t$$

具体的 AHT 算法的步骤如下:

3.11 Monge-Ampere 方程的数值解

日后拓展。

3.12 Semi-discrete OT 求解

本质上, 上文介绍的各种方法都是离散-离散形式下的算法。本小节介绍连续-离散下的 OT(Semi-discrete OT) 问题求解。这一领域的一个大佬有巴黎萨克雷大学的B.Levy。当然，这里不展示具体的算法步骤，只是计算半离散 OT 的 dual 问题的目标函数的导数与 Hessian 矩阵，主要是导数推导。有了梯度和 Hessian 矩阵，就可以用任何一阶优化算法与二阶优化 (如 L-BFGS , 牛顿法) 来计算。Hessen 矩阵的推导似乎也没找到一个看起来可读性高的（就我的工科背景），如顾险峰老师的书有几何视角的证明。当然书的水平起点有些高在只看这一部分时讲，但其教学视频在只看这一部分时还是可接受的。关于梯度，他在 2021 年的最优传输讲座 6有一个几何视角的梯度推导，他在 2020 年的教学视频里对 Hessian 矩阵有一个直观的几何性推导，容易接受!

- 梯度 $\nabla F(\psi)$ 推导

$\nabla F = \nabla\left(G(\sigma^{\psi},\psi) + \sum_{j=1}^{n}\psi_j\nu_j\right)$, 看第一项:

$$\begin{aligned}\frac{\partial G}{\partial \psi_j}(\psi) &= \lim_{t\to 0}\frac{G(\psi + te_j) - G(\psi)}{t}\\ &= \lim_{t\to 0}\frac{1}{t}\left\{\int_X \underbrace{\inf\left[c(x,y_1)-\psi_1,\ldots,c(x,y_j)-\psi_j-t,\ldots,c(x,y_n)-\psi_n\right]}_{①}\right.\\ &\left.-\underbrace{\inf_i\left[c(x,y_i)-\psi_i\right]}_{②}u(x)dx\right\}\end{aligned}$$

设 $x \in \mathrm{L}ag_{\psi}^{c}(y_m)$, 且 x 不在其边界上 (边界构成 0 测集可去掉), 有:

$② = c(x,y_m) - \psi_m$

$$① = \begin{cases} c(x,y_m)-\psi_m, m\neq j\\ c(x,y_j)-\psi_j - t, m = j\end{cases}$$

因此 ①-② $= \begin{cases} 0, m\neq j\\ -t, m = j\end{cases}$

对全部 $x \in X$ 进行积分, 只有 $x \in \mathrm{L}ag_{\psi}^{c}(y_j)$ 对积分有贡献, 因此有:

$$\frac{\partial G}{\partial \psi_j}(\psi) = -\int_{\mathrm{L}ag_{\psi}^{c}(y_j)} u(x)dx = -\int_{\mathrm{L}ag_{\psi}^{c}(y_j)} d\mu$$

因此:

$$\frac{\partial F}{\partial \psi_j}F(\psi) = \nu_j - \int_{\mathrm{L}ag_{\psi}^{c}(y_j)} d\mu$$

对应的对偶问题的最优性一阶条件 $\nabla F(\psi) = 0$ 为:

$$\nu_j = \int_{\mathrm{L}ag_{\psi}^{c}(y_j)} d\mu, \forall j \in \{1,2,\cdots,n\}$$

- Hessian$\nabla^2 F(\psi)$ 矩阵推导当 $c(x,y) = \frac{1}{2}(x-y)^2$ 时, Hessian 矩阵为:

$$\begin{aligned}\frac{\partial^2 F}{\partial\psi_i\partial\psi_j}(\psi) = \frac{\partial^2 G}{\partial\psi_i\partial\psi_j}(\psi) &= \int_{\mathrm{L}ag_{\psi}^{c}(y_i)\cap\mathrm{L}ag_{\psi}^{c}(y_j)}\frac{u(x)}{\|y_i - y_j\|}dS(x) \quad (i\neq j),\\ \frac{\partial^2 F}{\partial\psi_j^2}(\psi) = \frac{\partial^2 G}{\partial\psi_j^2}(\psi) &= -\sum_{i\neq j}\frac{\partial^2 G}{\partial\psi_i\partial\psi_j}(\psi).\end{aligned}$$

对于上述公式, 在此给一个直观的几何性证明: 仅考虑设 $u(x)$ 为常数的情况, 即 X 上为均匀分布。记即 y_i 关联的 *Laguerre cell* 面积为 ω_i 则 $\frac{\partial G}{\partial\psi_i}(\psi) = -\int_{\mathrm{L}ag_{\psi}^{c}(y_i)} u(x)dx = -\omega_i$, 因此 $\frac{\partial^2 G}{\partial\psi_i\partial\psi_j}(\psi) = -\frac{\partial\omega_i}{\partial\psi_j}$, 即考虑 ψ_j 变化对 ω_i 的影响。

显然若两个 *Laguerre cell* 不相交（几何上不相邻），如 $\mathrm{Lag}_{\psi}^{c}(y_k)$ 与 $\mathrm{Lag}_{\psi}^{c}(y_i)$，则 ψ_k 变化对 ω_i 产生影响，因此 $\frac{\partial^2 G}{\partial \psi_i \partial \psi_k} = 0$. 只有与 $\mathrm{Lag}_{\psi}^{c}(y_i)$ 相邻的 j 才会因为 ψ_j 的变化影响 ω_i.

如图4所示, 设 $\psi_j \to \psi_j + \delta\psi_j$,$\delta\psi_j > 0$, 则 $\mathrm{Lag}_{\psi}^{c}(y_j)$ 对应的面积 ω_j 变大，会导致 ω_i 变小。

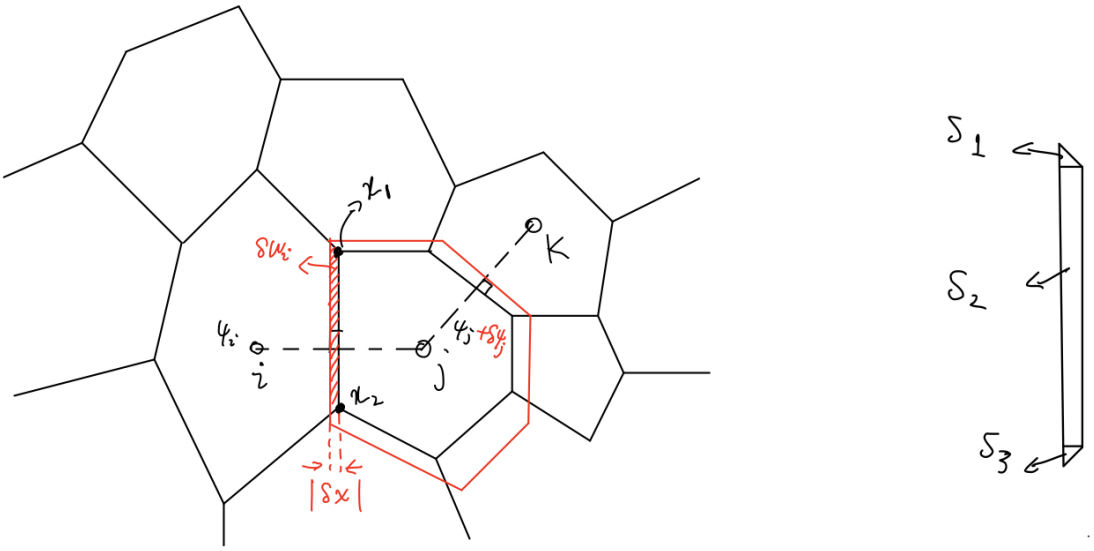


图 4: $c(x,y) = \frac{1}{2}(x-y)^2$ 下的半离散最优传输 Hessian 矩阵推导几何示意图

一方面，考虑 $\mathrm{Lag}_{\psi}^{c}(y_i)$ $\mathrm{Lag}_{\psi}^{c}(y_j)$ 交线处方程为

$$\frac{1}{2}(x-y_i)^2 - \psi_i = \frac{1}{2}(x-y_j)^2 - \psi_j$$

化简可得

$$-xy_i + \psi_i + y_i^2 = -xy_j + \psi_j + y_j^2$$

因此有:

$$\delta\psi_j = |\delta x| \cdot \|y_i - y_j\|$$

另一方面, $\delta\omega_i = -(S_1 + S_2 + S_3)$. 当 $\delta\psi_j \to 0$ 时，$\delta\omega_i \cong -S_2 = -|\delta x| \cdot \|x_1 - x_2\|$, 其中 $\|x_1 - x_2\|$ 为 $\mathrm{Lag}_{\psi}^{c}(y_i), \mathrm{Lag}_{\psi}^{c}(y_j)$ 交线的长度。

综合上述分析：对 $i \neq j$ 且 $\mathrm{Lag}_{\psi}^{c}(y_i), \mathrm{Lag}_{\psi}^{c}(y_j)$ 相交，交线为 $\|x_1 - x_2\|$ 有

$$\frac{\partial^2 G}{\partial \psi_i \partial \psi_j} = -\frac{\partial \omega_i}{\partial \psi_j} = -\frac{-|\delta x| \cdot \|x_1 - x_2\|}{|\delta x| \cdot \|y_i - y_j\|} = \frac{\|x_1 - x_2\|}{\|y_i - y_j\|}$$

上文中的 $\int_{Lag_\psi^c(y_i)\cap Lag_\psi^c(y_j)} u(x)dS(x) \quad (i \neq j)$ 部分相当于求交线（二维为交线，三维为面，更高维为超平面）的长度（面积/质量）。

最后，考虑 ψ_j 变化对面积 ω_j 本身的影响，可以完全等价于 ψ_j 对 j 的所有邻居面积 ω_i 的影响之和的负值。

有了 Hessian 矩阵、梯度表达式，可以使用如图5所示的 Newton 法求解半离散最优传输问题。

Input: a mesh that supports the source density u
the points $(y_j)_{j=1}^n$
the prescribed quantities $(\nu_j)_{j=1}^n$

Output: the (unique) Laguerre diagram Lag_ψ^c such that:
$$\int_{Lag_\psi^c(y_j)} u(x)dx = \nu_j \quad \forall j$$

---

(1) $\psi \leftarrow [0 \dots 0]$
(2) While convergence is not reached
(3) Compute ∇F and $\nabla^2 F$
(4) Find $p \in \mathbb{R}^n$ such that $\nabla^2 F(\psi)p = -\nabla F(\psi)$
(5) Find the descent parameter α
(6) $\psi \leftarrow \psi + \alpha p$
(7) End while

图 5: 文献 [2] 中基于 Newton 法的半离散最优传输算法

4 Gromov Wasserstein Distance

4.1 Gromov Wasserstein Distance 的定义

GW 距离弥补了 W 距离只能作用于同一度量空间下的限制，允许源域与目标域空间不同。其缺陷为：其目标函数二次非凸，很难应用到尺度比较大的背景下（如大规模的点云匹配）

$$\mathcal{GW}(\mathbf{P}_1,\mathbf{P}_2)=\inf_{T\in\mathbb{R}^{n_1\times n_2}}\sum_{i,j,i',j'}L(D(i,i'),\bar{D}(j,j'))T_{i,j}T_{i',j'}$$
$$\text{s.t.}\quad T\geq 0$$
$$\sum_j T_{i,j}=\omega_i,\forall i$$
$$\sum_i T_{i,j}=\omega_j,\forall j$$

其中，$D,\bar{D}$ 分别为源域、目标域上的成对距离矩阵，L 为用于度量两空间成对距离间的函数，如可取二范数，KL 散度。

记 $\mathcal{L}(D,\bar{D})\overset{\text{def.}}{=}(L(D_{i,i'},\bar{D}_{j,j'}))_{i,j,i',j'}$ 为 4 维张量，定义 $\mathcal{L}\otimes T\overset{\text{def.}}{=}\left(\sum_{i',j'}\mathcal{L}_{i,j,i',j'}T_{j,j'}\right)_{i,j}$ 因此，目标函数可简记为为：$f(T)=\langle\mathcal{L}(D_1,D_2)\otimes T,T\rangle$，可得梯度为 $\nabla f=2\mathcal{L}(D_1,D_2)\otimes T$。特别的，若上述距离函数 $L(x,y)$ 可分解为 $L(x,y)=f_1(x)+f_2(y)-h_1(x)h_2(y)$ 的形式，则:

$$\mathcal{L}(D,\bar{D})\otimes T=c_{D,\bar{D}}-h_1(D)Th_2(\bar{D})^\top$$

其中 $c_{D,\bar{D}}\overset{\text{def.}}{=}f_1(C)p\mathbb{1}_{N_2}^\top+\mathbb{1}_{N_1}q^\top f_2(\bar{C})^\top$ 与 T 无关!

例如 $L(x,y)=(x-y)^2=x^2+y^2-\sqrt{2}x\sqrt{2}y$ 时，$\mathcal{L}(D,\bar{D})\otimes T$ 形式如下：

$$\mathcal{L}(D,\bar{D})\otimes T=D^{\odot 2}\boldsymbol{p}\mathbf{1}_{n_2}^\top+\mathbf{1}_{n_1}\boldsymbol{q}^\top\bar{D}^{\odot 2}-2DT\bar{D}$$

在此情况下，优化问题简化 (用上约束条件，把 $<\mathcal{L}(D,\bar{D})\otimes T,T>$ 的前两项的 T 消去，变为与 T 无关的常数项) 为

$$\mathcal{GW}(\mathbf{P}_1,\mathbf{P}_2)=\inf_{T\in\Pi[\boldsymbol{p},\boldsymbol{q}]}-2<DT\bar{D},T>_F=\inf_{T\in\Pi[\boldsymbol{p},\boldsymbol{q}]}-2\mathrm{Tr}(DT\bar{D}T^{\mathbf{T}})=\inf_{T\in\Pi[\boldsymbol{p},\boldsymbol{q}]}-2\mathrm{Tr}(T^{\mathbf{T}}DT\bar{D})$$

Note:$<A,B>_F=\mathrm{Tr}(A^TB)=\mathrm{Tr}(B^TA)=\mathrm{Tr}(AB^T)=\mathrm{Tr}(BA^T)$

4.2 Mirror Descent Algorithm

4.2.1 一般性原理

对于无约束优化问题：

$$\min f(x)$$

梯度下降算法的迭代过程如下：

$$\mathbf{x}_{k+1}=\mathbf{x}_k-\alpha\nabla f(\mathbf{x}_k)$$

等价于求解如下最优化问题：

$$\begin{aligned}\mathbf{x}_{k+1} = &\quad \underset{\mathbf{x}}{\operatorname{argmin}}\, Q(\mathbf{x}, \mathbf{x}_k) \\ = &\quad \underset{\mathbf{x}}{\operatorname{argmin}}\, f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x} - \mathbf{x}_k \rangle + \frac{1}{2\alpha}\|\mathbf{x} - \mathbf{x}_k\|_2^2\end{aligned}$$

镜像梯度下降就是将上述 L2 范数 $\frac{\|\cdot\|^2}{2}$ 替换为更一般的 Bregman 散度：

$$\mathbf{x}_{k+1} = \underset{\mathbf{x}}{\operatorname{argmin}} f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x} - \mathbf{x}_k \rangle + \frac{1}{\alpha} B_\phi(\mathbf{x}, \mathbf{x}_k)$$

去掉常数项，则

$$\mathbf{x}_{k+1} = \underset{\mathbf{x}}{\operatorname{argmin}} \langle \nabla f(\mathbf{x}_k), \mathbf{x} \rangle + \frac{1}{\alpha} B_\phi(\mathbf{x}, \mathbf{x}_k)$$

带入 $B_\phi(x, y) = \phi(x) - \phi(y) - < \nabla\phi(y), x - y >$ 进一步化简可得：

$$\mathbf{x}_{k+1} = \underset{\mathbf{x}}{\operatorname{argmin}} \langle \nabla f(\mathbf{x}_k), \mathbf{x} \rangle + \frac{\phi(\mathbf{x}) - \langle \nabla\phi(\mathbf{x}_k), \mathbf{x} \rangle}{\alpha}$$

整理一下有：

$$\mathbf{x}_{k+1} = \underset{\mathbf{x}}{\operatorname{argmin}} \underbrace{\langle \nabla f(\mathbf{x}_k) - \frac{1}{\alpha}\nabla\phi(\mathbf{x}_k), \mathbf{x} \rangle + \frac{1}{\alpha}\phi(\mathbf{x})}_{Q(\mathbf{x}, \mathbf{x}_k)}$$

令 $\nabla_x Q = 0$, 有 $\nabla f(\mathbf{x}_k) - \frac{1}{\alpha}\nabla\phi(\mathbf{x}_k) + \frac{1}{\alpha}\nabla\phi(\mathbf{x}_{k+1}) = \mathbf{0}$, 整理得

$$\nabla\phi(\mathbf{x}_{k+1}) = \nabla\phi(\mathbf{x}_k) - \alpha\nabla f(\mathbf{x}_k)$$

因而：$\mathbf{x}_{k+1} = (\nabla\phi)^{-1}(\nabla\phi(\mathbf{x}_k) - \alpha\nabla f(\mathbf{x}_k))$.

考虑到勒让德对偶函数的性质——导函数互逆，即 $(\nabla\phi)^{-1} = \nabla(\phi^*)$, 其中 ϕ^* 为 ϕ 的 Legendre 对偶。因此上述表达式也可以写成：

$$\mathbf{x}_{k+1} = \nabla(\phi^*)(\nabla\phi(\mathbf{x}_k) - \alpha\nabla f(\mathbf{x}_k))$$

对于有约束的优化问题:

$$\min_{x \in \mathcal{X}} f(x)$$

对一般梯度下降加一个投影步，形成投影梯度下降：

$$\mathbf{x}^{k+1} = P_{\mathcal{X}}(\mathbf{x}^k - \alpha f'(\mathbf{x}^k))$$

其中投影的本质为居于选定的距离函数，在可行解集中选出离目标最近的可行解。常采用的距离函数为欧式距离，其投影步的公式如下：

$$P_{\mathcal{X}}(\mathbf{y}) = \underset{\mathbf{x} \in \mathcal{X}}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$$

对于有约束优化问题的（投影）镜像梯度下降，其投影步中的距离函数为 Bregman 散度：

$$P_{\mathcal{X},\phi}(\mathbf{y}) = \underset{\mathbf{x}\in\mathcal{X}}{\operatorname{argmin}} B_\phi(\mathbf{x},\mathbf{y})$$

综合梯度下降步与投影步，可得有约束优化问题的镜像下降算法的迭代公式：

$$\mathbf{x}_{k+1} = P_{\mathcal{X},\phi}[(\nabla\phi)^{-1}(\nabla\phi(\mathbf{x}_k) - \alpha\nabla f(\mathbf{x}_k))]$$

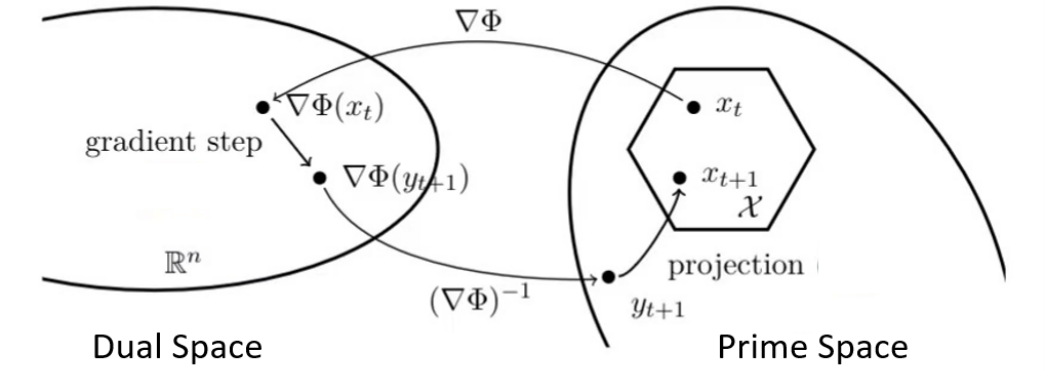


图 6: Mirror Descent Method 示意图

Note: 对比一般的梯度下降更新式，上述镜像梯度下降迭代式反应的过程为: 将原空间 ($\mathcal{X}$) 先变到对偶空间 ($\nabla\phi(\mathcal{X})$)，在对偶空间进行更新 ($\nabla\phi(\mathbf{x}_k) - \alpha\nabla f(\mathbf{x}_k)$)，然后再次对偶回原空间 ($\mathcal{X}$), 此为镜像之谓也。图6很好的诠释了镜像下降算法

对于少数约束、选择合适的 ϕ 可以获得约束优化问题迭代步的解析形式！下面给出 $\phi = x\log(x)$ 下，镜像下降算法中下降步的具体形式。（投影步需根据约束推导，如 Sec.3.1.4和 Sec.3.4）

由于 $\nabla\phi = 1 + \log(x)$, 因此 $(\nabla\phi)^{-1} = e^{x-1}$, 因此镜像下降步为：

$$\begin{aligned}\mathbf{x}_{k+1} = (\nabla\phi)^{-1}(\nabla\phi(\mathbf{x}_k) - \alpha\nabla f(\mathbf{x}_k)) &= e^{\nabla\phi(\mathbf{x}_k)-\alpha\nabla f(\mathbf{x}_k)-1} \\ &= e^{\log(\mathbf{x}_k)-\alpha\nabla f(\mathbf{x}_k)} \\ &= \mathbf{x}_k \cdot e^{-\alpha\nabla f(\mathbf{x}_k)}\end{aligned}$$

4.2.2 Shinkhorn 迭代-镜像梯度算法

取 Bregman 散度中 $\phi(x) = x\log x$，利用镜像梯度算法，可得传输方案的迭代公式如下：

$$T \leftarrow \mathrm{Proj}^{\mathrm{KL}}_{\Pi[\boldsymbol{p},\boldsymbol{q}]}\left(T \odot e^{-\tau(\nabla f(T) - \epsilon\nabla\mathbf{H}(T))}\right) = \mathrm{Proj}^{\mathrm{KL}}_{\Pi[\boldsymbol{p},\boldsymbol{q}]}\left(T \odot e^{-\tau(2\mathcal{L}\otimes T + \epsilon\log(T)}\right)$$

其中 τ 为更新的步长，ϵ 为熵正则的正则项系数。

根据 Sec3.2.1介绍的 KL 散度与熵正则 Kantorovich 最优传输的关系可知，投影步正好为一熵正则 Kantorovich 问题，则迭代式变为：

$$T \leftarrow \underset{\boldsymbol{\pi}\in\Pi[\boldsymbol{p},\boldsymbol{q}]}{\operatorname{argmin}} < -\epsilon_1 \log(T \odot e^{-\tau(2\mathcal{L}\otimes T+\epsilon\log(T)}), \boldsymbol{\pi} > -\epsilon_1 \mathbf{H}(\boldsymbol{\pi})$$

记 $\bar{C} = -\epsilon_1 \log(T \odot e^{-\tau(2\mathcal{L}\otimes T+\epsilon\log(T)})$, 化简得：

$$\begin{aligned}\bar{C} &= -\epsilon_1 \log(T \odot e^{-\tau(2\mathcal{L}\otimes T+\epsilon\log(T)}) \\ &= -\epsilon_1 \log(T) + \epsilon_1\tau(2\mathcal{L}\otimes T + \epsilon\log(T))\end{aligned}$$

若 $\epsilon\tau = 1$, 取 $\epsilon_1 = \epsilon$, 则

$$\bar{C} = 2\mathcal{L}\otimes T$$

最终更新的迭代式变为：

$$T \leftarrow \underset{\boldsymbol{\pi}\in\Pi[\boldsymbol{p},\boldsymbol{q}]}{\operatorname{argmin}} < 2\mathcal{L}\otimes T, \boldsymbol{\pi} > -\epsilon\mathbf{H}(\boldsymbol{\pi})$$

即在上述参数假定下，采用镜像梯度算法将熵正则的 Gromov-Wasserstein 问题转化为一个熵正则的 Wasserstein 问题，可通过 Sinkhorn 算法计算

Remark: 由于要求 $\tau\epsilon = 1$, 导致正则项和步长矛盾、冲突，最终使得该算法在求解原 Gromov-Wasserstein 问题时会遭受的收敛性问题。当然也可以不要 $\tau\epsilon = 1$ 这个设定，此时 $\bar{C} = 2\tau\epsilon\mathcal{L}\otimes T - \epsilon\log(T)(1-\epsilon\tau)$。

4.2.3 Bregman 迭代-镜像梯度算法

根据镜像梯度算法梯度步的解析形式以及 Sec.3.4中投影项的解析形式, 可以轻松的写出 Bregman 迭代-镜像梯度算法的迭代式。

对于 Gromow-Wasserstein 问题，梯度就是 $2\mathcal{L}\otimes T$.

但你会发现与文献 [3]、[4] 中梯度不一样。那是因为文献对目标函数用上了约束，化简了目标函数，丢弃了常数项。正如前面定义那一节那样，优化的目标变为：

$$\mathcal{GW}(\mathbf{P}_1,\mathbf{P}_2) = \inf_{T\in\Pi[\boldsymbol{p},\boldsymbol{q}]} -2 < DT\bar{D}, T >_F = \inf_{T\in\Pi[\boldsymbol{p},\boldsymbol{q}]} -2\mathrm{Tr}(DT\bar{D}T^{\mathbf{T}}) = \inf_{T\in\Pi[\boldsymbol{p},\boldsymbol{q}]} -2\mathrm{Tr}(T^{\mathbf{T}}DT\bar{D})$$

因此，梯度变为 $-4DT\bar{D}$, 这就是文献[4] 所使用的。因此其完成的迭代式为：

$$\begin{aligned}\pi^{(k)} &\leftarrow \boldsymbol{\pi}^{(k)} \odot \exp\left(\gamma(4\alpha D\pi\bar{D} - (1-\alpha)M_{12})\right) \\ \pi^{(k)} &\leftarrow \mathrm{diag}(\boldsymbol{p}./\boldsymbol{\pi}^{(k)}\mathbb{1}_m)\boldsymbol{\pi}^{(k)} \\ \pi^{(k)} &\leftarrow \boldsymbol{\pi}^{(k)} \odot \exp\left(\gamma(4\alpha D\pi\bar{D} - (1-\alpha)M_{12})\right) \\ \pi^{(k)} &\leftarrow \boldsymbol{\pi}^{(k)}\mathrm{diag}(\boldsymbol{q}/\boldsymbol{\pi}^{(k)\top}\mathbb{1}_n)\end{aligned}$$

其中，γ 为步长，α 为 Fusion GW 中的组合因子

而文献[3]，直接把上述 $\mathcal{GW}(\mathbf{P}_1, \mathbf{P}_2)$ 的常数项去掉了，并且按照 Bregman 迭代的想法把目标函数里的两个 T, 分成了传输映射 π 和传输映射 ω, 这样对任一个传输映射求导，变成了 $DT\bar{D}$,T 取 π 或 ω. 因此其完整的迭代式为:

$$\pi \leftarrow \pi \odot \exp(\alpha D\pi\bar{D}), \quad \pi \leftarrow \mathrm{diag}(\boldsymbol{p}./\pi\mathbb{1}_m)\pi,$$
$$\pi \leftarrow \pi \odot \exp(\alpha D\pi\bar{D}), \quad \pi \leftarrow \pi\mathrm{diag}(\boldsymbol{q}./\pi^T\mathbb{1}_n)$$

其中 α 为步长。

4.3 Proximal Point Algorithm

见 Sec.5.4

5 Fusion Gromov Wasserstein Distance

5.1 Fusion Gromov Wasserstein Distance 的定义

$$\mathcal{FGW}(\mathbf{P}_1, \mathbf{P}_2) = \inf_{T\in\mathbb{R}^{n_1\times n_2}} \left[(1-\alpha)\langle T, \mathbf{M}_{1,2}\rangle_F + \alpha \sum_{i,j,i',j'} \mathcal{L}(D_1(i,i'), D_2(j,j'))T_{i,j}T_{i',j'}\right]$$
$$\text{s.t.} \quad T \geq 0$$
$$\sum_j T_{i,j} = \omega_i, \forall i$$
$$\sum_i T_{i,j} = \omega_j, \forall j$$
$$\text{s.t.} \quad T\cdot\mathbf{1} = \omega_1; \quad T^\top\cdot\mathbf{1} = \omega_2; \quad T \geq 0$$

或简写为:

$$\mathcal{FGW}(\mathbf{P}_1, \mathbf{P}_2) = \inf_{T\in\Pi[\boldsymbol{p},\boldsymbol{q}]} (1-\alpha) < M_{12}, T > +\alpha < \mathcal{L}\otimes T, T >$$

记目标函数 $f(T) = (1-\alpha) < M_{12}, T > +\alpha < \mathcal{L}\otimes T, T >$, 则对 T 的梯度 $\nabla f(T) = (1-\alpha)M_{12} + 2\alpha\mathcal{L}\otimes T$。

其中，$D, \bar{D}$ 分别为源域、目标域上的成对距离矩阵，M_{12} 表示源域、目标域之间的距离矩阵,L 为用于度量两空间成对距离间的函数，如可取二范数，KL 散度。

5.2 Bellman 迭代-镜像梯度算法

见 Sec.4.2.3.

5.3 The Frank–Wolfe/Conditional Gradient Method

Frank–Wolfe/Conditional Gradient Method 为一类可行方向法，其思想上构造可行下降方向，然后在此可行下降方向上走特定的步长，使得目标函数值下降。而构造可行方向的方式基于线性化!

- **Frank–Wolfe 算法的原理**

  对于如下优化问题:

$$\begin{aligned} \min \quad & f(\boldsymbol{x}) \\ s.t. \quad & A_1\boldsymbol{x} \geq \boldsymbol{b}^1 \\ & A_2\boldsymbol{x} = \boldsymbol{b}^2 \end{aligned}$$

  记上述约束对应的可行解集为 Ω. 对于第 $k+1$ 次迭代，在 $x^k \in \Omega$ 处将目标函数线性展开:

$$f(x) \approx f(x^k) + \nabla f(x^k)^T(x - x^k) = \nabla f(x^k)^T x + [f(x^k) - \nabla f(x^k)^T x^k]$$

  用之代替目标函数，等价于求解如下优化问题:

$$\begin{aligned} \min \quad & \nabla f(x^k)^T x \\ s.t. \quad & A_1 x \geq b^1 \\ & A_2 x = b^2 \end{aligned}$$

  设其最优解为 y^k. 定义 $d^k = y^k - x^k$. 显然由于 y^k 为最优解则 $\nabla f(x^k)d^k = \nabla f(x^k)(y^k - x^k) \leq 0$.

  定理: 设 $x^k \in \Omega$,y^k 为上述线性化后的最优极点,$d^k = y^k - x^k$,$z^k = \nabla f(x^k)d^k$ 则:

  (1) 若 $z^k < 0$, 则 d^k 为原问题在 x^k 处的可行下降方向;

  (2) 若 $z^k = 0$, 则 x^k 为原问题的 KKT 点。

  证明: 若 $z^k < 0$, 则 $\nabla f(x^k)d^k < 0$，即 d^k 为下降方向。$\forall \lambda \in [0,1]$,$x^k + \lambda d^k = (1-\lambda)x^k + \lambda y^k \in \Omega$(由于 Ω 为凸集)，因此 d^k 为可行下降方向。

若 $z^k=0$,$z_k=\nabla f(x^k)^T d^k=0\Rightarrow \nabla f(x^k)^T x^k=\nabla f(x^k)^T y^k\Rightarrow x^k$ 为线性化后优化问题的最优解和 KKT 点。则 $\exists u$、ν, 使 $\begin{cases}\nabla f(x^k)-A_1^T\boldsymbol{u}-A_2^T\nu=\boldsymbol{0}\\ u^T(A_1\boldsymbol{x}^k-\boldsymbol{b}^1)=0\\ \boldsymbol{u}\geq 0\end{cases}$，这正好亦为原问题的 KKT 条件，即 x^k 为原问题的 KKT 点。 ■

当目标函数连续可微，可行域有界时，可以证明 Frank–Wolfe 算法是全局收敛于原问题的 KKT 点。

- **条件梯度算法的一般框架**

(1) 选定初始点 $x^1\in\Omega,\varepsilon,k=1$.

(2) 构造搜索方向。计算线性化后的最优化问题，求解最优点 $y^k\in\Omega$. 构造可行下降方向 $d^k=y^k-x^k$, 求出 $z^k=\nabla f(x^k)^\top d^k$.

(3) 若 $z^k=0$，结束，x^k 为最优解，否则求解最优步长：$\arg\min_{0\leq\lambda\leq 1} f(x^k+\lambda d^k)$.

(4) 获得新迭代点，$x^{k+1}=x^k+\lambda_k d^k, k\Leftarrow k+1$, 转（2）.

特别地，若目标函数为严格凸二次函数，则最优步长为

$$\lambda_k=\min\left\{-\frac{z_k}{(d^k)^T Hd^k},1\right\},H=\nabla^2 f(x)$$

- **求解 Fusion Gromov Wasserstein 问题的 Frank–Wolfe 算法**

根据上述条件梯度法的基本框架，求解 Fusion Gromov Wasserstein 问题共两个主要步骤。

首先为线性化。可以求得目标函数的梯度为 $G(T)=(1-\alpha)M_{AB}^q+2\alpha L(C_1,C_2)^q\otimes T$. 因此线性化后为：

$$\bar{T}^k=\arg\min_{T\in\Pi[\boldsymbol{p},\boldsymbol{q}]}<G(T^k),T>$$

此时可以将 Fusion Gromov Wasserstein 问题，转为一个成本函数为 $G(T^k)$ 的 Kantorovich 问题，可以采用 Sinkhorn 算法求解或者网络单纯形法。

然后就是构造可行下降的方向 $d^k=\bar{T}^k-T^k$, 判断是否已最优。否则，进行一维步长搜索 $\lambda^k=\arg\min_{\lambda\in[0,1]} f(T^k+\lambda d^k)$，求出新的迭代点 $T^{k+1}=T^k+\lambda^k d^k$. 由于 FGW 本身就是个二次函数，因此求解最优搜索步长是可解析的！

下面我们取 $L(x,y)=\|x-y\|_2^2$ 为例进行推导！此时有：

$$\mathcal{L}(D,\bar{D})\otimes T=c_{D,\bar{D}}-2DT(\bar{D})^\top$$

$$f(T^k+\lambda d^k)=(1-\alpha)<M_{12},T^k+\lambda d^k>+\alpha<c_{D,\bar{D}}-2D(T^k+\lambda d^k)(\bar{D})^\top,T^k+\lambda d^k>$$
$$=\lambda<(1-\alpha)M_{12}+\alpha c_{D,\bar{D}},d^k>-2\alpha\underbrace{<D(T^k+\lambda d^k)\bar{D}^\top,T^k+\lambda d^k>}_{①}$$
$$+(1-\alpha)<M_{12},T^k>+\alpha<c_{D,\bar{D}},T^k>$$
$$①=\lambda^2<D(d^k)\bar{D}^\top,d^k>+\lambda<D(T^k)\bar{D}^\top,d^k>+\lambda<D(d^k)\bar{D}^\top,T^k>+<D(T^k)\bar{D}^\top,T^k>$$

因此:

$$f(T^k+\lambda d^k)=-2\alpha\lambda^2<D(d^k)\bar{D}^\top,d^k>-\lambda<D(T^k)\bar{D}^\top,d^k>-2\alpha\lambda<D(d^k)\bar{D}^\top,T^k>$$
$$-2\alpha\lambda<D(d^k)\bar{D}^\top,T^k>+\lambda<(1-\alpha)M_{12}+\alpha c_{D,\bar{D}},d^k>+FGW(T^k)$$

写成 $f(\lambda)=a\lambda^2+b\lambda+c$ 形式。对比有:

$$\begin{cases}a=-2\alpha<D(d^k)\bar{D}^\top,d^k>\\ b=<(1-\alpha)M_{12}+\alpha c_{D,\bar{D}},d^k>-2\alpha<D(T^k)\bar{D}^\top,d^k>-2\alpha<D(d^k)\bar{D}^\top,T^k>\\ c=FGW(T^k)\end{cases}$$

现在就是简单的二次函数 $f(\lambda)=a\lambda^2+b\lambda+c$ 在区间 $[0,1]$ 取极小值问题了。利用初中知识，分情况讨论:

1. $a>0$:
   * $-\frac{b}{2a}\geq 1$: $\lambda^*=1$
   * $-\frac{b}{2a}\leq 0$: $\lambda^*=0$
   * $0<-\frac{b}{2a}<1$: $\lambda^*=-\frac{b}{2a}$

   综上有: $\lambda^*=\min(1,\max(0,\frac{-b}{2a}))$
2. $a=0$:
   * $b>0$: $\lambda^*=0$
   * $b<0$: $\lambda^*=1$
3. $a<0$:
   * $-\frac{b}{2a}\geq\frac{1}{2}$: $\lambda^*=0,\Rightarrow b+a\geq 0$
   * $-\frac{b}{2a}\leq\frac{1}{2}$: $\lambda^*=1,\Rightarrow b+a\leq 0$

   可对 $a=0,a<0$ 情况何并，有:
   * $b+a\leq 0$: $\lambda^*=1$

* $b + a \geq 0$: $\lambda^* = 0$

根据上述讨论求出 λ^*, 因此下一步迭代为 $T^{k+1} = T^k + \lambda^* d^k = \lambda^* \bar{T}^k + (1 - \lambda^*) T^k$.

5.4 Proximal Point Algorithm

使用近端点法优化的问题为:

$$T^{k+1} = \underset{T \in \Pi[\boldsymbol{p}, \boldsymbol{q}]}{\operatorname{argmin}} (1 - \alpha) < M_{12}, T > + \alpha < \mathcal{L} \otimes T, T > + \gamma \mathrm{KL}(T | T^k)$$

其中 $\mathrm{KL}(T|T^k) = \sum_{i,j} T \log \frac{T_{i,j}}{T^k_{i,j}} - T_{i,j} + T^k_{i,j}$.

利用镜像下降方法来得到上述问题等价于求解如下熵正则 Kantorovich 问题:

$$T^{k+1} = \underset{T \in \Pi[\boldsymbol{p}, \boldsymbol{q}]}{\operatorname{argmin}} < (1 - \alpha) M_{12} + 2\alpha \mathcal{L} \otimes T^k - \gamma \log(T^k), T > - \gamma \mathrm{H}(T)$$

其中 $\mathrm{H}(T) = -\sum_{i,j} T_{ij} (\log(T_i j) - 1)$.

证明: 由于 $\nabla f = (1 - \alpha) M_{12} + 2\alpha (\mathcal{L} \otimes T) + \gamma \log(\frac{T}{T^k})$, 则 $\nabla f(T^k) = (1 - \alpha) M_{12} + 2\alpha (\mathcal{L} \otimes T^k)$.

利用镜像下降算法，可得更新步为:

$$T^{k+1} = \underset{T \in \Pi[\boldsymbol{p}, \boldsymbol{q}]}{\operatorname{argmin}} \mathrm{KL}(T^k \odot e^{-\tau((1-\alpha) M_{12} + 2\alpha (\mathcal{L} \otimes T^k))})$$

令 $K = T^k \odot e^{-\tau((1-\alpha) M_{12} + 2\alpha (\mathcal{L} \otimes T^k)) = e^{-\bar{C}/\gamma}}$, 则:

$$\begin{aligned} \bar{C} &= -\gamma \log K = -\gamma (\log(T^k) - \tau((1 - \alpha) M_{12} + 2\alpha \mathcal{L} \otimes T^k)) \\ &= \gamma \tau ((1 - \alpha) M_{12} + 2\alpha \mathcal{L} \otimes T^k) - \gamma \log(T^k) \end{aligned}$$

其中 τ 为步长，γ 为正则项系数。取 $\tau\gamma = 1$ 时可得上述结果。 ■

当然你可以不用 $\tau\gamma = 1$ 这个假设，获得更一般的结果。

另外，也可以把 GW 项的 $\mathcal{L} \otimes T$ 换成 $\mathcal{L} \otimes T^k$, 把原优化问题展开重组也会获得上述结论。

6 OT 的其他拓展形式

本节简短描述 OT 的扩展，比如 sliced Wasserstein distance, Quantized Gromov-Wasserstein, unbalance OT, 表述其定义式，说明其用途。

6.1 Partition OT

6.1.1 简介与定义式

最优传输要求质量守恒，即满足 $\mu_1(T^{-1}(B)) = \mu_2(B)$, 但是对于颜色、形状匹配等某些应用，这个要求有些严格：被匹配的两个对象（分布）可能有任意的质量，或者只需要传输总质量的一部分。解决方案包含本小节的 Partition OT 和下一小节的 Unbalanced OT。国内研究者有中科大的陈世炳教授,菲尔兹将得主 A. Figalli 也研究过此问题。

Partition OT 研究两个概率分布间传输给定部分质量 $0 \leq s \leq \min(\|\boldsymbol{p}\|_1, \|\boldsymbol{q}\|_1)$，并使传输成本最低。此时，可行的传输方案集变为 $\Pi^p(\boldsymbol{p}, \boldsymbol{q}) = \{\boldsymbol{T} \in \mathbb{R}_+^{|\boldsymbol{p}|\times|\boldsymbol{q}|} \mid \boldsymbol{T}\mathbb{1}_{|\boldsymbol{q}|} \leq \boldsymbol{p}, \boldsymbol{T}^\mathsf{T}\mathbb{1}_{|\boldsymbol{p}|} \leq \boldsymbol{q}, \mathbb{1}_{|\boldsymbol{p}|}^\mathsf{T}\boldsymbol{T}\mathbb{1}_{|\boldsymbol{q}|} = s\}$. Partition Warsserstein distance 的定义如下

$$\mathcal{PW}(\boldsymbol{p}, \boldsymbol{q}) = \min_{\boldsymbol{T}\in\Pi^p(\boldsymbol{p},\boldsymbol{q})} < \boldsymbol{C}, \boldsymbol{T} >_F$$

类似地，相比于 Gromov Wasserstein，Partition GW 的差异也就仅在约束上。

6.1.2 求解原理

求解最优传输大多需要放松约束，也就是边缘分布那一项。文献[5] 提出了不同的思路，那就是引入 “Virtual or dummy” 点（其可视为具有任意特征的点）来解决传输时质量不相同问题或者部分传输问题，这个策略在求图编辑距离时也有应用。可以证明，通过在源域和目标域中各加入一个虚空点，可以把 Patition OT 化为一个相对标准的 Wasserstein 问题。而求解 Partition GW 距离文献 [5] 提出 Frank-Wolfe 算法（又称投影梯度算法，与上文解 Fusion GW 的思路一致），不过每个线性化步因为约束的缘故为 Partition Wasserstein 问题，因此用解 Partition Wasserstein 问题的方法。下面着重描述如何通过加 dummy 点将 Partition Wasserstein 问题拓展成一般的 Wasserstein 问题。

定义： $\bar{\boldsymbol{C}} = \begin{bmatrix} \boldsymbol{C} & \xi\mathbb{1}_{|\boldsymbol{p}|} \\ \xi\mathbb{1}_{|\boldsymbol{q}|}^\mathsf{T} & 2\xi + A \end{bmatrix}$, $n = |\boldsymbol{p}|, m = |\boldsymbol{q}|$, $\bar{\boldsymbol{p}} = [\boldsymbol{p}, \|\boldsymbol{q}\|_1 - s], \bar{\boldsymbol{q}} = [\boldsymbol{q}, \|\boldsymbol{p}\|_1 - s]$.

命题： *若 $A > \max(C_{ij})$ 且标量 ξ 有界，则*

$$\mathcal{W}(\bar{\boldsymbol{p}}, \bar{\boldsymbol{q}}) - \mathcal{PW}(\boldsymbol{p}, \boldsymbol{q}) = \xi(\|\boldsymbol{p}\|_1 + \|\boldsymbol{q}\|_1 - 2s)$$

且扩展的 Wasserstein 问题的最优传输方案 $\bar{\boldsymbol{T}}^*$ 去掉最后一行最后一列为对应 Partition Wasserstein 问题的最优传输方案 $\boldsymbol{T}^*$. 证明见文献 [2] 附录.

6.2 Unbalanced OT

同样聚焦于 OT 传输两个相同的总质量的缺陷，但有别于 Partition OT 传一个固定的部分质量，unbalanced OT [6] 通过加惩罚项来松弛总质量守恒的约束，允许源域的输出质量不等于其包含的质量，也允许目标域的输入质量不等于其目标质量。其问题定义如下

$$\mathcal{W}^u(\boldsymbol{p},\boldsymbol{q}) = \min_{\boldsymbol{T}\in\mathbb{R}_+^{|\boldsymbol{p}|\times|\boldsymbol{q}|}} < C, \boldsymbol{T} >_F + \tau_1 \mathbf{D}_\phi(\boldsymbol{T}\mathbb{1}_{|\boldsymbol{q}|}\|\boldsymbol{p}) + \tau_2 \mathbf{D}_\phi(\boldsymbol{T}^\mathsf{T}\mathbb{1}_{|\boldsymbol{p}|}\|\boldsymbol{q})$$

其中 (τ_1,τ_2) 用于控制允许的质量变化，当 $\tau_1,\tau_2 \to +\infty$ 时，类似罚函数，迫使传输过程质量守恒，退化为经典（平衡）最优传输，当仅其中一项趋于 $+\infty$ 是为 semi-unbalance OT。

当 $\mathbf{D}_\phi = KL$ 散度时，如果再加入熵正则项：

$$\min_{\boldsymbol{T}\in\mathbb{R}_+^{|\boldsymbol{p}|\times|\boldsymbol{q}|}} < C, \boldsymbol{T} >_F + \tau_1 \mathbf{KL}(\boldsymbol{T}\mathbb{1}_{|\boldsymbol{q}|}\|\boldsymbol{p}) + \tau_2 \mathbf{KL}(\boldsymbol{T}^\mathsf{T}\mathbb{1}_{|\boldsymbol{p}|}\|\boldsymbol{q}) - \varepsilon \mathbf{H}(T)$$

则亦可通过 Sinkhorn 算法快速计算，其中变动的项为：

$$\boldsymbol{u}^{(\ell+1)} \overset{\text{def.}}{=} \left(\frac{\boldsymbol{p}}{\mathbf{K}\boldsymbol{v}^{(\ell)}}\right)^{\frac{\tau_1}{\tau_1+\varepsilon}} \quad \text{and} \quad \boldsymbol{v}^{(\ell+1)} \overset{\text{def.}}{=} \left(\frac{\boldsymbol{q}}{\mathbf{K}^\mathrm{T}\boldsymbol{u}^{(\ell+1)}}\right)^{\frac{\tau_2}{\tau_2+\varepsilon}}$$

6.3 Sliced OT 与 Randon Transform

Wasserstein 距离计算是昂贵的，特别是对于高维问题，以及大尺度问题。虽然有熵正则方法进行加速计算，但在某些情况下，问题本身有很好的结构，如两个一维的概率密度之间的 Wasserstein 距离有闭式解，并且可以高效的估计。这个属性，促进了 Sliced OT 的发展。sliced-Wasserstein distance 故名思意，他是对高维概率密度进行无限多次的线性投影，获得很多一维分布表示（类似于对高维分布做线性分片,slice），然后计算这些一维分布间的 Wasserstein 距离的平均值。把高维问题，拆分为多个简单的一维问题，它相对于经典的 Wasserstein 距离，计算更简单，计算需求更少。特别地，这里的线性投影与 Radon tranform 关联密切，后者是 CT 的关键技术。当然了，把高维投影到一维，会有一些信息丢失，要想充分反应高维，只能增大投影的数量。在高维的背景下，数据常常在一些薄流形（Manifold），因此要想捕获这种数据分布的结构所需选择的线性投影数量将会增长的很快。因此，减少所需线性投影的数量是改进 Sliced Wasserstein distance 计算的关键，一些方案包含 max-sliced-Wasserstein, 它找单个线性投影，使得在该投影空间的距离最大；还有引入线性子空间投影；引入 Monte Carlo 方法等。文献 [7], 基于线性投影与 Radon 变换的联系，利用 *generalized* Radon 变换获得 generalized

sliced Wasserstein. 并且将线性投影变成非线性，在某些条件下，仍然是一个度量，并且引入非线性可以更好建模复杂（高维）分布的结构，减少所需投影的数量。

根据文献 [7]，设两个一维概率分布为 $\boldsymbol{p}, \boldsymbol{q}$，其累计分布函数（CDF）为 $F_\mu(x) = \mu((-\infty, x]) = \int_{-\infty}^{x} p(\tau)d\tau$ 和 $F_\nu(x) = \nu((-\infty, x]) = \int_{-\infty}^{x} q(\tau)d\tau$，则他们间的最优传输映射为 $f(x) = F_\nu^{-1}(F_\mu(x))$ 他们的 W 距离为

$$W_p(\mu, \nu) = (\int_X d^p(x, F_\nu^{-1}(F_\mu(x))))^{\frac{1}{p}} = (\int_0^1 d^p(F_\mu^{-1}(z), F_\nu^{-1}(z))dz)^{\frac{1}{p}}, F_\mu(x) = z$$

用 Radon Transform 对高维分布做线性投影，然后计算投影后的一维概率分布的平均即得 Sliced Wasserstein:

$$SW_p(\boldsymbol{p}, \boldsymbol{q}) = (\int_{\mathbb{S}^{d-1}} W_p^p(\mathcal{R}\boldsymbol{p}(., \theta), \mathcal{R}\boldsymbol{q}(., \theta))d\theta)^{\frac{1}{p}}$$

$$\mathcal{R}f(t, \theta) = \int_{\mathbb{R}^d} f(x)\delta(t - < x, \theta >)dx, (t, \theta) \in \mathbb{R} \times \mathbb{S}^{d-1}$$

其中 Radon 变换 $\mathcal{R} : L^1(\mathbb{R}^d) \to L^1(\mathbb{R} \times \mathbb{S}^{d-1})$, $L^1(\mathbb{R}) = \{f : \mathbb{R}^d \to \mathbb{R} \mid \int_{\mathbb{R}^d} |f(x)|dx < \infty\}, \mathbb{S}^{d-1} \subset \mathbb{R}^d$ 为单位球面. 求解上述 SW 距离可以采用 Monte-Carlo 方法，把积分换成求和后平均。但是对高维球面采样总会出现一些看起来奇怪的事（交大许志钦老师的可解释深度学习也有提及相关的高维空间一些“反常识”的事情）：在高维球面采样的点总与给定向量正交，这会导致 $W_p(\mathcal{R}\boldsymbol{p}(., \theta), \mathcal{R}\boldsymbol{q}(., \theta)) \approx 0$ 以很高的概率出现，因此要有选择的采样 θ. 另一简便做法为 maximum sliced p-Wasserstein 距离：

$$\max -SW_p(\boldsymbol{p}, \boldsymbol{q}) = \max_{\theta \in \mathbb{S}^{d-1}} W_p(\mathcal{R}\boldsymbol{p}(., \theta), \mathcal{R}\boldsymbol{q}(., \theta))$$

Radon 变换的几何意义？为何 Radon 变换为线性投影？

Radon 变换相当于 $f(\boldsymbol{x})$ 在一族超平面（hyperplane）$H(t, \theta) = \{x \in \mathbb{R}^d \mid < x, \theta >= t\}$ 上做积分。固定 θ，$\mathcal{R}I(t, \theta) : \mathbb{R} \to \mathbb{R}, t \mapsto \mathbb{R}$, 并且这个积分是在超平面组 $H(t, \theta) = \{x \in \mathbb{R}^d \mid < x, \theta >= t, t \in \mathbb{R}\}$ 上做，该超平面族的法向量就是 θ，相当于高维函数 f 向 θ 的方向做投影。向一个方向的投影等价于对与该方向的正交的其他方向做积分。如二维概率密度函数向 y 轴投影，是通过对 x 做积分获得 $f_Y(y) = \int_{-\infty}^{+\infty} f(x, y)dx$。对应地，$\theta$ 就是这里的 y，$H(., \theta)$ 就是这里的被积区域 $x \in (-\infty, +\infty)$. 更直观一点，向 y 投影，那么与 y 方向正交的超平面就是 $\{(x, y) \in \mathbb{R}^2 \mid x = t, t \in \mathbb{R}\}$, 被积区域就是这族超平面上的点.

如果将 Randon 变换里的 $< x, \theta >$ 换成更一般的 $g(x, \theta)$, 得到 Generalized Radon Transform:

$$\mathcal{G}f(t, \theta) = \int_{\mathbb{R}^d} f(x)\delta(t - g(x, \theta)dx$$

, 那么就发展得到了文献[7] 提出的 Generalized Sliced-Wasserstein Distance 以及对应的 maximum generalized sliced Wasserstein distance:

$$GSW_p(\boldsymbol{p},\boldsymbol{q}) = (\int_{\Omega_\theta} W_p^p(\mathcal{G}\boldsymbol{p}(.,\theta),\mathcal{G}\boldsymbol{q}(.,\theta))d\theta)^{\frac{1}{p}}$$
$$\max -GSW_p(\boldsymbol{p},\boldsymbol{q}) = \max_{\theta\in\Omega_\theta} W_p(\mathcal{G}\boldsymbol{p}(.,\theta),\mathcal{G}\boldsymbol{q}(.,\theta))$$

其中 Ω_θ 为 $g(.,\theta)$ 可行解的紧集, 如 $g(x,\theta) =< x,\theta >$ 时, $\Omega_\theta = \mathbb{S}^{d-1}$. 需要指出, 这个函数 g 要满足一些很好的性质, 如 Hessian 矩阵严格正定, 对 θ 为齐次 $(g(t,\lambda\theta) = \lambda(t,\theta))$, 光滑性（$C^\infty$）. 具体到计算，就是用 MC 方法，看起来也比较简单，并且里面的经 Generalized Radon Transform 得到的一维分布间的 Wasserstein distance 也有近似的解析公式。函数 g 的具体条件、数值计算公式的具体细节见文献[7]。

6.4 Quantized Gromov-Wasserstein

Gromov-Wasserstein 问题, 对不大可数值计算的 Gromov Hausdoff 距离进行了 relax, 本质是做非凸优化，为两个度量空间找一个最佳的匹配，但是对大规模数据上因计算问题受限，如大规模点云场景。一些解决方案通常对两个空间进行划分，然后选出各子空间的代表元进行全局最优传输（或匹配），然后递归地在全局配合好的子空间对内部进行最优传输。如上文的 Sliced OT, 其中的划分就是那个线性投影。文献[8] 集成了这种划分的范式 (or 分而治之策略 divide-and-conquer strategy), 引入量化到 Gromov-Wasserstein 问题, 提出 Quantized GW 用于解决划分后子空间之间匹配问题，最终可将求解规模由 1∼10K 扩大到 100K∼1M.

一些定义如下：$(X, d_X\ \mu_X)$ 为有限度量空间，X 的一个划分为 $\{U^1, U^2, \cdots, U^m\}$ 要求 $\forall i, U^i \neq \emptyset$ 且 $\forall i,j, U^i \cap U^j = \emptyset$. 记 $x^p \in U^p$ 为子集 U^p 的代表元，记 $\mathcal{P}_X = \{(x^1,U^1),(x^2,U^2),\cdots,(x^m,U^m)\}$ 为 *m-pointed partition*,$X^m = \{x^1,x^2,\cdots,x^m\}$ 为划分代表元形成的集合。可以赋予 X^m, 视为 X 的一个商空间，一个（自然）测度 $\mu_{\mathcal{P}_X}$, 对 μ_X 通过那个划分映射 *pushforward* 得到，即 $\mu_{\mathcal{P}_X}(B) = \mu_X(P^{-1}(B)), B \subset X^m,\quad P^{-1}(B) = \cup_{i\in\{p:x^p\in B\}} U^i$。保留距离度量 d_X 定义在此代表元空间为 $d_{X|X^m}$. 将 $X^m = (X^m, d_{X|X^m}, \mu_{\mathcal{P}_X})$ 作为原空间 X 的量化表示。同样，还可以定义子空间 $(U^p, d_{X|U^p}, \mu_{U^p})$, 其中 $d_{X|U^p}$ 定义同 d_X 只不过作用空间约束在子集 U^p,$\mu_{U^p} = (\mu_X(U^p))^{-1}\mu_{X|U^p} = \frac{\mu_{X|U^p}}{\mu_X(U^p)}$, $\mu_{X|U^p}$ 定义同 μ_X 只不过作用在子集 U^p 上。这样，可以对原空间 X 利用子空间测度诱导一个新的测度 $\bar{\mu}_{U^p}(A) := \mu_{U^p}(A\cap U^p), \forall A \subset X$，简单视为条件分布。

记 $\mu_{m,n} \in \mathcal{C}(\mu_{\mathcal{P}_X}, \mu_{\mathcal{P}_Y})$ 为两个商空间之间可行的 *coupling*，可以理解为 $\mathcal{P}_X$ $\mathcal{P}_Y$ 间传输的联合概率密度。记 $\mu_{x^p,y^q} \in \mathcal{C}(\mu_{U^p}, \mu_{V^q})$ 为两个子空间的可行的 *coupling*，同样可

理解为 U^p, V^q 间的联合概率密度，同上的方式可以用这个子空间诱导出整个空间在子空间约束下的 coupling。$\bar{\mu}_{x^p,y^q} \in \mathcal{C}(\bar{\mu}_{U^p}, \bar{\mu}_{V^q})$，简单地可视为条件分布。$\bar{\mu}_{x^p,y^q}(A,B) = \mu_{x^p,y^q}(A \cap U^p, B \cap V^q), \forall A \subset X, B \subset Y$.

根据全概然公式有

$$\mu(x,y) = \sum_{p,q} \mu_{m,n}(x^p, y^q)\bar{\mu}_{x^p,y^q}(x,y)$$

其中 $\mu(x,y)$ 为度量空间 X, Y 之间的 *quantization coupling*。文献 [8] 证明了 *quantization coupling* 就是 X, Y 空间之间的 *coupling*，即是一个可行解。因此求解 Quantized Gromov-Wasserstein 问题是求解 Gromov-Wasserstein 的一个上界。

Quantized Gromov-Wasserstein 计算思路如下：

1. **数据划分：** 图数据可采用 *networkx* 的 *Fluid community detection algorithm* 进行图划分，然后对每个划分应用一些指标，如 maximal PageRank、中心度等指标选择代表元。点云可采用均匀采样得到代表元，然后对采样点进行 Voronoi 划分得到划分。或者可以采用诸如 K-means、谱聚类等各类聚类算法进行划分、选出代表元;

2. **全局匹配（$\mu_{m,n}(x^p, y^q)$）：** 对两个空间 X，Y 的量化（划分）空间 X^m, Y^n 求解 Gromov-Wesserstein 问题, 找到最优传输方案（划分空间的联合概率分布）;

3. **局部匹配（μ_{x^p,y^q}）：** 对每个 $x^p \in X^m$,$y^q \in Y^n$, 解下列 OT 问题，

$$\min_{\mu_{x^p,y^q} \in \mathcal{C}(\mu_{U^p}, \mu_{V^q})} \sum_{x \in \mu_{U^p}, y \in \mu_{V^q}} (d_X(x, x^p) - d_Y(y - y^q))^2 \mu_{x^p,y^q}(x,y)$$

   计算每个成对的子空间下的最优传输（联合概率密度）得到的局部最优传输方案，就是以全局匹配为条件的局部分配概率分布;

4. **生成最终传输方案：** 用全概然公式计算原问题 X, Y 空间间的联合概率密度，即最优传输方案。

$$\mu(x,y) = \sum_{p,q} \mu_{m,n}(x^p, y^q)\bar{\mu}_{x^p,y^q}(x,y)$$

6.5 Multi-Marginal OT

我们上面求的 Kantorovich 问题中传输方案的只有两个边缘（分布），一个自然的想法是能不能把“二维变高维”，把两个边缘变成多个边缘。这时我们求解的传输

方案就不是二维的矩阵 $\mathbf{P} \in \mathbb{R}^{n_1 \times n_2}$，而是高维的张量 $\mathbf{P} \in \mathbb{R}^{n_1 \times \ldots \times n_S}$. 设有 S 个分布 $\mathbf{a_s} \in \Sigma_{n_s}, s = 1, 2, \cdots S$. 则我们求解的可行解即为：

$$\mathbf{U}(\mathbf{a}_s)_s = \left\{ \mathbf{P} \in \mathbb{R}_{\geq 0}^{n_1 \times \ldots \times n_S} : \forall s = 1, 2, \cdots, S, \forall i_s = 1, \cdots n_s, \sum_{\ell \neq s} \sum_{i_\ell = 1}^{n_\ell} \mathbf{P}_{i_1, \ldots, i_S} = \mathbf{a}_{s, i_s} \right\}$$

优化问题为：

$$\min_{\mathbf{P} \in \mathbf{U}(\mathbf{a}_s)_s} \langle \mathbf{C}, \mathbf{P} \rangle \stackrel{\text{def.}}{=} \sum_{s=1}^{S} \cdots \sum_{i_s = 1}^{n_s} \mathbf{C}_{i_1, \ldots, i_S} \mathbf{P}_{i_1, \ldots, i_S}$$

做 Entropic regulazation，有：

$$\min_{\mathbf{P} \in \mathbf{U}(\mathbf{a}_s)_s} \langle \mathbf{C}, \mathbf{P} \rangle - \varepsilon \mathbf{H}(\mathbf{P})$$

我们可以很自然的把之前的 Sinkhorn 算法推广到这个场景下：

记多重指标 $i = (i_1, \cdots, i_S)$, 则 $\mathbf{P}_i^* = \mathbf{K}_i \prod_{s=1}^{S} \mathbf{u}_{s, i_s}$.

其中 $\mathbf{K} \stackrel{\text{def.}}{=} e^{-\frac{\mathbf{C}}{\varepsilon}}$,$\mathbf{u}_s \in \mathbb{R}_+^{n_s}$ 为那个 Scaling Vector, 其更新通过 Sinhorn 算法：依次（循环）遍历 $s = 1, 2, \cdots, S$ 至收敛:(公式分母说明：求和是对除第 s 维固定外全体多重指标，注意 K 的指标与各个 u 的指标都是对应好的)

$$\mathbf{u}_{s, i_s} \leftarrow \frac{\mathbf{a}_{s, i_s}}{\sum_{l \neq s} \sum_{i_l = 1}^{n_l} \mathbf{K}_i \prod_{r \neq s} \mathbf{u}_{r, i_r}}$$

显然在上述内循环的更新中，可以利用已更新的 $\mathbf{u}_s$ 的结果去更新之后的 $\mathbf{u}_{s+k}, k > 0$.

当然，也可以直接用 Bregman 迭代3.4. 为了方便表示，我们把一个张量沿第 s 维求和（与 numpy 求和中的 axis、pytorch 的 dim，最终求和后张量的维度为去掉（沿此方向求和得标量）指定的 axis or dim 不同）记为：

$$[Sum_s(\mathbf{P})]_i := \sum_{j_1, \cdots, j_{s-1}, j_{s+1}, \cdots, j_S} \mathbf{P}_{j_1, \cdots, j_{s-1}, i, j_{s+1}, \cdots, j_S}, i = 1, \cdots, n_s$$

这里为只保留此维度方向。比如形状为 (3,4,5) 的张量沿 s=3（下标从 1 开始）求和为一个 5 维向量。在此记号下，可行域为：

$$\mathbf{U}(\mathbf{a}_s)_s = \left\{ \mathbf{P} \in \mathbb{R}_{\geq 0}^{n_1 \times \ldots \times n_S} : Sum_s(\mathbf{P}) = \mathbf{a}_s, \forall s = 1, \cdots, S \right\}$$

因而，采用 IBP 方法为，依次从 $s = 1, 2, \cdots, S$, 更新

$$\forall j = (j_1, \cdots, j_S), \ \mathbf{P}_j \leftarrow \frac{[\mathbf{a}_s]_{j_k}}{[Sum_s(\mathbf{P})]_{j_k}} \mathbf{P}_j$$

7 Numerical Method for Wasserstein Barycenter

7.1 OT type Wasserstein Barycenter

7.2 GW type Wasserstein Barycenter

7.3 Computing Wasserstein Barycenter via Input Convex Neural Networks

本节参考：Scalable Computations of Wasserstein Barycenter via Input Convex Neural Networks

8 Wasserstein Space and Calculus

8.1 Wasserstein 度量空间

我们在如下的测度空间上研究：$\mathscr{P}_p(\Omega) := \left\{\mu \in \mathscr{P}(\Omega) : \int_\Omega |x|^p \mathrm{d}\mu < +\infty\right\}$，其中 $p \geq 1$

对于 $\mu, \nu \in \mathscr{P}_p(\Omega)$, 定义距离：

$$W_p(\mu,\nu) := \min\left\{\int_{\Omega\times\Omega} |x-y|^p \mathrm{d}\gamma : \gamma \in \Pi(\mu,\nu)\right\}^{\frac{1}{p}}$$

此测度空间要求积分有限的原因为: 由于 $\|x-y\|^p \leq C(\|x\|^p + \|y\|^p)$, 因此可以保证 $\mathcal{W}_p(\mu,\nu) < \infty$ 亦有限。上述 $C = 2^{p-1}$，这完全利用 Lp 范数的三角不等式 (以 z=0 为媒介，$\|x-y\|_p < \|x\|_p + \|y\|_p$) 和凸函数性质 (平均的值小于等于值的平均)。

之后简记 $\left[\int_{\Omega\times\Omega} \|x-y\|^p \mathrm{d}\gamma\right]^{1/p}$ 为 $||x-y||_{L^q(\gamma)}$.

对于 $1 \leq p \leq q$, 我们根据范数的性质 $\|f\|_p \leq \|f\|_q$, 有：

$$\left(\int_\Omega |x-y|^p \mathrm{d}\gamma\right)^{\frac{1}{p}} = ||x-y||_{L^p(\gamma)} \leq ||x-y||_{L^q(\gamma)} = \left(\int_\Omega |x-y|^q \mathrm{d}\gamma\right)^{1/q}$$

设 m 为测度，根据 Holder 不等式, 有：

$$\begin{aligned}
&\int_\Omega |f|^p \,\mathrm{d}m = \int_\Omega |f|^p \cdot 1 \mathrm{d}m \\
&\leq \left(\int_\Omega |f|^{p\cdot q/p} \mathrm{d}m\right)^{p/q} \left(\int_\Omega |1|^{q/(q-p)} \mathrm{d}m\right)^{(q-p)/q} \\
&= \left(\int_\Omega |f|^q \mathrm{d}m\right)^{p/q} (m(\Omega))^{(q-p)/q}.
\end{aligned}$$

即：$\|f\|_{L^p} \leqslant (m(\Omega))^{1/p-1/q}\|f\|_{L^q}$. 特别的，当 m 为概率测度，$m(\Omega)=1$, 即得上述大控小。

相反的不等式“仍然”成立，$\dim(\Omega)=\sup\{|x-y|:x,y\in\Omega\}$,then

$$\left(\int \|x-y\|^p \mathrm{d}\gamma\right)^{\frac{1}{p}} \leq \dim(\Omega)^{\frac{p-1}{p}} \left(\int \|x-y\|\mathrm{d}\gamma\right)^{\frac{1}{p}}$$

即 $W_p(\mu,\nu)\leq CW_1(\mu,\nu)^{\frac{1}{p}}, C=\dim(\Omega)^{(p-1)/p}$

命题（$W_p(\cdot,\cdot)$ 为距离）： p-Wasserstein 距离为定义在 $\mathscr{P}(\Omega)$ 上的度量。i.e.,$\forall\mu,\nu,\varrho\in\mathscr{P}(\Omega)$ 满足：

$$1.W_p(\mu,\nu)\geq 0$$

$$2.W_p(\mu,\nu)=W_p(\nu,\mu)$$

$$3.W_p(\mu,\nu)=0 \quad \textit{iff} \quad \mu=\nu$$

$$4.W_p(\mu,\nu)\leq W_p(\mu,\rho)+W_p(\rho,\nu)$$

（1），（2），（3）完全可利用定义化为 L^p 范数的性质来看，下面只证明 W_p 满足三角不等式。

证明:(从输出映射角度)

$$W_p(\mu,\nu)\leq \left(\int |S(\mathrm{T}(x))-x|^p \mathrm{d}\mu(x)\right)^{\frac{1}{p}} = ||S\circ \mathrm{T}-\mathrm{id}||_{L^p(\mu)}$$

$$\leq ||S\circ \mathrm{T}-\mathrm{T}||_{L^p(\mu)} + ||\mathrm{T}-\mathrm{id}||_{L^p(\mu)}.$$

又有（参数变换）：

$$||S\circ\mathrm{T}-\mathrm{T}||_{L^p(\mu)} = \left(\int |S(\mathrm{T}(x))-\mathrm{T}(x)|^p \mathrm{d}\mu(x)\right)^{\frac{1}{p}} = \left(\int |S(y)-y|^p \mathrm{d}\varrho(y)\right)^{\frac{1}{p}} = ||S-\mathrm{id}||_{L^p(\varrho)}$$

只要取 T 为 $\mu\to\varrho$, S 为 $\varrho\to\nu$ 的最优传输映射, 即可获得三角不等式：$W_p(\mu,\nu)\leq W_p(\mu,\varrho)+W_p(\varrho,\nu)$ ■

证明:(利用 L^p 空间的三角不等式)

$$\begin{aligned} W_p(\mu,\nu) &\leq (\mathbb{E}\|X-Y\|^p)^{1/p} \\ &= (\mathbb{E}\|X-Z+Z-Y\|^p)^{1/p} \\ &\leq (\mathbb{E}\|X-Z\|^p)^{1/p} + (\mathbb{E}\|Z-Y\|^p)^{1/p} \end{aligned}$$

只要取 $Z\sim\varrho$, 且 (X,Z) (Z,Y) 正好为 $\mu\to\varrho,\varrho\to\nu$ 的最优传输方案，即可获得三角不等式：$W_p(\mu,\nu)\leq W_p(\mu,\varrho)+W_p(\varrho,\nu)$ ■

定义（$\mathbb{W}_p$ 空间）： 给定 *Polish* 空间 Ω（可分的完备度量空间, 换言之, 一个 Hausdoff 拓扑空间，其上定义了与其拓扑相容的距离，且该空间在此距离定义下完备，也就是任何柯西列都有收敛子列）,$p\geq 1$。定义 p 阶 Wasserstein 空间为，$\mathbb{W}_p(\Omega)=(\mathscr{P}_p(\Omega),W_p)$。

命题: $p \geq 1$,$(\mathscr{P}_p(\mathbb{R}^n), W_p)$ 为完备度量空间。

8.2 Otto's Calculus

本节将介绍 Otto 等人对 $\mathbb{W}_p$ 空间定义的一些黎曼几何结构，包含范数、内积、梯度。关于 $\mathbb{W}_p$ 上的恒速测地线，请查看 Sec.2.6.2。参考 Figalli 的小册子，我们从 Benamou-Brenier 公式

$$W_2^2(\mu_0,\mu_1) = \inf_{v_t,\nu_t}\left\{\int_0^1\left(\int_\Omega |v_t|^2\mu_t dxdt\right)\Big|(\mu_t,v_t)_{t\in[0,1]} solves \partial_t\mu_t + \mathrm{div}(\mu_t v_t) = 0, \mu_t v_t\cdot\mathbf{n}|_{\partial\Omega} = 0\right\}.$$

开始进行自然的引入。以下所涉及的 μ_t 都视为概率密度。

在之前的 Sec.2.6.2中，我们类比了一般黎曼几何中的测地距离，定义了 $\partial_t\mu_t$ 的 *Wasserstein norm*:

$$\|\partial_t\mu_t\|_{\mu_t}^2 := \inf_{v_t}\left\{\int_\Omega |v_t|^2\mu_t dx\Big|\ \mathrm{div}(v_t\mu_t) = -\partial_t\mu_t, \mu_t v_t\cdot\mathbf{n}|_{\partial\Omega} = 0\right\}$$

将 W_p 可写成了如下形式：$W_2^2(\bar\mu_0,\bar\mu_1) = \inf_{\mu_t}\left\{\int_0^1 \|\partial_t\mu_t\|_{\mu_t}^2\, dt\Big|\mu_0 = \bar\mu_0, \mu_1 = \bar\mu_1\right\}$.

随后我们将根据 (μ_t, v_t) 满足的约束，改写 $\partial_t\mu_t$ 的 *Wasserstein norm* $\|\partial_t\mu_t\|_{\mu_t}$.

设 v_t 为上述 $\|\partial_t\mu_t\|_{\mu_t}^2$ 对应的最优解。给此最小值加扰动，$v_t + \varepsilon\frac{\omega}{\mu_t}$, 且要求 ω 无散，即 $\mathrm{div}(\omega) = 0$。此时 $\mathrm{div}\Big(\Big(v_t + \varepsilon\frac{w}{\rho_t}\Big)\nu_t\Big) = -\partial_t\mu_t$, 即 $v_t + \varepsilon\frac{\omega}{\mu_t}$ 仍为可行解。由于 v_t 为问题的最优解，有:

$$\begin{aligned}\int_\Omega |v_t|^2\mu_t dx &\leq \int_\Omega \Big|v_t + \varepsilon\frac{w}{\rho_t}\Big|^2\mu_t dx\\ &= \int_\Omega |v_t|^2\mu_t dx + 2\varepsilon\int_\Omega \langle v_t, w\rangle dx + \varepsilon^2\int_\Omega \frac{|w|^2}{\mu_t}dx.\end{aligned}$$

右侧在 $\varepsilon = 0$ 时取得最小值，根据最优的一阶条件有：$\int_\Omega < v_t, \omega > dx = 0$. 即最优解 v_t 与任意无散场正交。根据 Helmholtz decomposition 定理，任意向量场可以分解为无散场和势能场的和，因此 v_t 必为某个势能函数诱导的势能场：$v_t = \nabla\Psi_t$。此时，连续方程与无 flux 边界条件（也就是约束）变为:

$$\begin{cases}\mathrm{div}(\mu_t\nabla\Psi_t) = -\partial_t\mu_t & \mathrm{in}\Omega,\\ \frac{\partial\Psi_t}{\partial\mathbf{n}} = 0 & \mathrm{on}\partial\Omega.\end{cases}$$

对应的 $\|\partial_t\mu_t\|_{\mu_t}^2 = \int_\Omega |\nabla\Psi_t|^2\mu_t dt$，满足如上约束。由此，我们可以进行推广：给定概率密度 $\rho \in \mathscr{P}_2(\Omega)$,$h : \Omega \to \mathbb{R}$, 且 $\int_\Omega h dx = 0$, 定义 h 在给定的 ρ 下的 *Wasserstein norm* 为

$$\|h\|_\rho^2 := \int_\Omega |\nabla\psi|^2\rho dx, \quad \text{where} \quad \begin{cases}\mathrm{div}(\rho\nabla\psi) = -h & \mathrm{in}\Omega,\\ \frac{\partial\psi}{\partial\mathbf{n}} = 0 & \mathrm{on}\partial\Omega.\end{cases}$$

Remark: 这里要求 $\int_\Omega h dx = 0$ 保证了椭圆 PDE 的可解性: $\int_\Omega h dx = -\int_\Omega \mathrm{div}(\rho\nabla\psi)dx = -\int_{\partial\Omega}\frac{\partial\psi}{\partial\mathbf{n}}\rho = 0$ (最后一个等式由边界条件获得)。或考虑 $\int_\Omega \partial_t\mu_t dx = \frac{d}{dt}\int_\Omega \mu_t dx = \frac{d}{dt}1 = 0$

有了范数的定义,自然可以推广到内积。给定 $h_1, h_2 : \Omega \to \mathbb{R}$, 且 $\int_\Omega h_1 dx = \int_\Omega h_2 dx = 0$。定义 h_1, h_2 在 ρ 下的 *Wasserstein scalar product* 为:

$$\langle h_1, h_2\rangle_\rho := \int_\Omega \nabla\psi_1\cdot\nabla\psi_2\rho dx, \quad \text{where} \quad \begin{cases} \mathrm{div}(\rho\nabla\psi_i) = -h_i & \text{in}\Omega, \\ \frac{\partial\psi_i}{\partial\mathbf{n}} = 0 & \text{on}\partial\Omega. \end{cases}$$

最后，我们定义 Wasserstein 空间下，泛函的梯度.

定义（$\mathbb{W}_p$ 下泛函梯度）: 给定泛函 $\mathcal{F} : \mathscr{P}_2(\Omega) \to \mathbb{R}\cup\{+\infty\}$, 其在 $\rho \in \mathscr{P}_2(\Omega)$ 处，相对于 *Wasserstein scalar product* 的梯度（如果存在的话）是唯一的函数 $\mathrm{grad}_{W_2}\mathcal{F}[\rho]$, 使得:

$$\left\langle \mathrm{grad}_{W_2}\mathcal{F}[\rho], \frac{\partial\rho_\varepsilon}{\partial\varepsilon}\Big|_{\varepsilon=0}\right\rangle_\rho = \frac{d}{d\varepsilon}\Big|_{\varepsilon=0}\mathcal{F}[\rho_\varepsilon]$$

对任意的光滑曲线 $\rho_\varepsilon : (-\varepsilon_0, \varepsilon_0) \to \mathscr{P}_2(\Omega)$， $\rho_0 = \rho$ 成立。

定义（$\mathbb{W}_p$ 下泛函的一阶变分 First Variations）: 沿用上述假设和条件，记泛函 $\mathcal{F}[\rho]$ 的一阶变分（如果存在）为 $\frac{\delta\mathcal{F}[\rho]}{\delta\rho}$, 使得

$$\frac{d}{d\varepsilon}\Big|_{\varepsilon=0}\mathcal{F}[\rho_\varepsilon] = \int_\Omega \frac{\delta\mathcal{F}[\rho]}{\delta\rho}(x)\frac{\partial\rho_\varepsilon(x)}{\partial\varepsilon}\Big|_{\varepsilon=0}dx$$

对任意的光滑曲线 $\rho_\varepsilon : (-\varepsilon_0, \varepsilon_0) \to \mathscr{P}_2(\Omega)$， $\rho_0 = \rho$ 成立。

Remark: 这里的 ρ 等都是概率密度。[1] 中采用概率测度进行的定义:

$$\frac{d}{d\varepsilon}\Big|_{\varepsilon=0}\mathcal{F}(\varrho + \varepsilon\chi) = \int \frac{\delta\mathcal{F}}{\delta\varrho}(\varrho)\mathrm{d}\chi$$

对任意测度 ϱ 的扰动 $\chi = \tilde{\varrho} - \varrho$ with $\tilde{\varrho} \in L_c^\infty(\Omega)\cap\mathscr{P}(\Omega)$ 都成立。

基于一阶变分的定义，则泛函梯度满足:

$$\left\langle \mathrm{grad}_{W_2}\mathcal{F}[\rho], \frac{\partial\rho_\varepsilon}{\partial\varepsilon}\Big|_{\varepsilon=0}\right\rangle_\rho = \int_\Omega \frac{\delta\mathcal{F}[\rho]}{\delta\rho}\frac{\partial\rho_\varepsilon}{\partial\varepsilon}\Big|_{\varepsilon=0}dx$$

注意 *Wasserstein scalar product* 是有约束条件的。一方面记 ψ 为约束:$\mathrm{div}(\nabla\psi\rho) = -\frac{\partial\rho_\varepsilon}{\partial\varepsilon}\Big|_{\varepsilon=0}$ 的解，且满足 Zero Neumann 边界条件 i.e.,$\frac{\partial\psi_t}{\partial\mathbf{n}} = 0$, 因此上式变为:

$$\left\langle \mathrm{grad}_{W_2}\mathcal{F}[\rho], \frac{\partial\rho_\varepsilon}{\partial\varepsilon}\Big|_{\varepsilon=0}\right\rangle_\rho = -\int_\Omega \frac{\delta\mathcal{F}[\rho]}{\delta\rho}\mathrm{div}(\nabla\psi\rho)dx = \int_\Omega \nabla\frac{\delta\mathcal{F}[\rho]}{\delta\rho}\cdot\nabla\psi\rho dx$$

另一方面，根据 *Wasserstein scalar product* 定义，对比获得 $\psi_1 = \frac{\delta\mathcal{F}[\rho]}{\delta\rho}$, 根据约束条件最终得:

$$\mathrm{grad}_{W_2}\mathcal{F}[\rho] = -\mathrm{div}\Big(\nabla\Big(\frac{\delta\mathcal{F}[\rho]}{\delta\rho}\Big)\rho\Big)$$

Example1 $\mathcal{F}[\rho]=\int_\Omega U(\rho(x))dx$, 其中 $U:\mathbb{R}\to\mathbb{R}$.

此时

$$\frac{d}{d\varepsilon}\Big|_{\varepsilon=0}\int_\Omega U(\rho_\varepsilon(x))dx=\int_\Omega U'(\rho(x))\frac{\partial\rho_\varepsilon(x)}{\partial\varepsilon}\Big|_{\varepsilon=0}dx$$

对比有：$\frac{\delta\mathcal{F}[\rho]}{\delta\rho}(x)=U'(\rho(x))$. 因而梯度为：

$$\mathrm{grad}_{W_2}\mathcal{F}[\rho]=-\mathrm{div}\big(\rho\nabla[U'(\rho)]\big)=-\mathrm{div}\big(\rho U''(\rho)\nabla\rho\big)$$

特别地，其 $U(x)=x\log x$, 则 $U''=\frac{1}{x}$, 则 $\mathrm{grad}_{W_2}\mathcal{F}[\rho]=-\Delta\rho$.

取 $U(s)=\frac{s^m}{m-1}, m\neq 1$，则 $\mathrm{grad}_{W_2}\mathcal{F}[\rho]=-\mathrm{div}\big(\rho m\rho^{m-2}\nabla\rho\big)=-\Delta(\rho^m)$

Example2 $\mathcal{F}[\rho]=\int_\Omega V(x)\rho dx$, 其中 $V:\Omega\to\mathbb{R}$.

则 $\frac{\delta\mathcal{F}[\rho]}{\delta\rho}(x)=V(x)$, 此时，$\mathrm{grad}_{W_2}\mathcal{F}[\rho]=-\mathrm{div}(\rho\nabla V)$

Example3 $\mathcal{F}[\rho]=\frac{1}{2}\iint h(x-y)\rho(x)\rho(y)dxdy$, 其中 $h:\mathbb{R}^n\to\mathbb{R}$ 且 $h(z)=h(-z)$.

$$\frac{\delta\mathcal{F}[\rho]}{\delta\rho}(x)=h*\rho(x)=\int_\Omega h(x-y)\rho(y)dy$$
$$\mathrm{grad}_{W_2}\mathcal{F}[\rho]=-\mathrm{div}\big((\nabla h*\rho)\rho\big)$$

最后，为了下一节保证 Wasserstein 梯度流的存在性与唯一性（如收敛到泛函最小点），我们在此引入 $\mathbb{W}_2$ 空间上泛函凸性的概念，这个概念最早由 McCann 引入。

定义（Displacement Convexity or W_2- 凸） 称泛函 $\mathcal{F}:\mathscr{P}_2(\Omega)\to\mathbb{R}$ 为 W_2- 凸或 Displacement Convexity，如果一维映射 $[0,1]\ni t\mapsto\mathcal{F}[\rho_t]$ 是凸的，对所有 $\mathbb{W}_2$ 空间的测地线 $\rho:[0,1]\to\mathscr{P}_2(\Omega)$ 都成立。

Note: 要求 $\mathbb{W}_2$ 上所有测地线都凸有点过于严格，因此有一个弱化版本。若对任意两个概率测度 $\mu,\nu\in\mathscr{P}_2$, 若连接两者的 $\mathbb{W}_2$ 测地线 (可能不唯一) 中存在一条为凸，则称泛函 $\mathcal{F}[\rho]$ 为弱位移凸性。特别地，若 $\mathcal{F}[\rho]=\int_\Omega U(\rho(x))dx$, 且满足 $U:[0,+\infty)\to\mathbb{R}$ 为凸的 Lipschitz 连续函数，$U(0)=0$, 则弱位移凸性等价于位移凸性，如 $U(\rho)=\rho\log(\rho)$.

我们接下来将探索上述条件下 ($\mathcal{F}[\rho]=\int_\Omega U(\rho(x))dx$, 且满足 $U:[0,+\infty)\to\mathbb{R}$ 为凸的 Lipschitz 连续函数，$U(0)=0$)，函数 $U(\rho(x))$ 还需满足什么条件，能保证泛函 $\mathcal{F}[\rho]$ 为位移凸。

正好 $\mathbb{W}_2$ 对应成本函数为 $c(x,y)=\frac{|x-y|^2}{2}$, 根据 Brenier 定理知，连接任意绝对连续测度 ρ_0,ρ_1 的最优传输映射 $T=\nabla(\frac{x^2}{2}-\phi)=\nabla u_\phi$/最优传输方案 $(id,T)_\sharp\rho_0$ 唯一，因此测地线唯一，具体形似为：$\rho_t=(T_t)\sharp\rho_0, T_t=id+t(T-id)$. 此时 T_t 亦将为 $\rho_0\to\rho_t$ 的最优传输映射，其将满足 Jacobi 方程：

$$\rho_t\circ T_t=\frac{\rho_0}{\det\nabla T_t},\nabla T_t=(1-t)id+t\nabla^2u_\phi$$

因此：

$$\begin{aligned}\mathcal{F}[\rho_t] &= \int_\Omega \frac{U(\rho_t(y))}{\rho_t(y)}\rho_t(y)dy \xlongequal[dy=\det(\nabla T_t)dx]{y=T_t(x)} \int_\Omega \frac{U(\rho_t\circ T_t)}{\rho_t\circ T_t}\rho_0 dx \\ &= \int_\Omega U\Big(\frac{\rho_0}{\det\nabla T_t}\Big)\det\nabla T_t dx \\ &= \int_\Omega U\Big(\frac{\rho_0}{\det((1-t)\mathrm{id}+t\nabla^2 u_\phi)}\Big)\det\Big((1-t)\mathrm{id}+t\nabla^2 u_\phi(x)\Big)dx\end{aligned}$$

由于 Brenier 势能函数为凸函数，则 $\nabla^2 u_\phi$ 半正定，则其特征值 $\lambda_1,\cdots,\lambda_d\geq 0$, 因此：

$$\begin{aligned}D(x,t) &:= \det\big((1-t)\mathrm{Id}+t\nabla^2\varphi(x)\big)^{1/d} \\ &= \det\begin{pmatrix}(1-t)+t\lambda_1(x) & \cdots & 0\\ 0 & \ddots & 0\\ 0 & \cdots & (1-t)+t\lambda_d(x)\end{pmatrix}^{1/d} \\ &= \prod_{i=1}^d\big((1-t)+t\lambda_i(x)\big)^{1/d}.\end{aligned}$$

则

$$\mathcal{F}[\rho_t] = \int_\Omega U\Big(\frac{\rho_0(x)}{D(x,t)^d}\Big)D(x,t)^d dx$$

可以证明 $t\mapsto D(x,t)$ 为凹函数 (见 [1]Page 271, 引理 7.26)。紧接着将有如下命题成立 (参考 [9]P90 的证明)

命题： 设 $U:[0,+\infty)\to\mathbb{R}$,$U(0)=0$，$\Omega$ 为凸集。若映射：

$$(0,\infty)\ni s\mapsto U\Big(\frac{1}{s^d}\Big)s^d$$

为凸函数且不增，则泛函 $\mathcal{F}[\rho]=\int_\Omega U(\rho(x))dx$ 是 W_2 凸的。

基于上述命题，则 U 取如下形式，

$$U(s) := \begin{cases} s\log(s) & & \Rightarrow & \partial_t\rho_t=\Delta\rho(\text{heat eq.}),\\ \frac{1}{m-1}s^m & \text{for} \quad m>1 & \Rightarrow & \partial_t\rho_t=\Delta(\rho^m)(\text{porous medium eq.}),\\ \frac{1}{m-1}s^m & \text{for} \quad m\in[1-\frac{1}{d},1) & \Rightarrow & \partial_t\rho_t=\Delta(\rho^m)(\text{fast diffusion eq.})\end{cases}$$

对应的泛函 $\mathcal{F}[\rho]=\int_\Omega U(\rho(x))dx$ 是 W_2 凸的, 其对应的 Wasserstein 梯度流方程形式如右侧所示。

Claim: 若 $V:\Omega\to\mathbb{R}$ 为凸函数，则 $\mathcal{F}[\rho]:=\int_\Omega V(x)\rho(x)dx$ 为 W_2 凸的。

若 $h:\mathbb{R}^d\to\mathbb{R}$ 为凸函数, 则泛函 $\mathcal{F}[\rho]:=\iint_{\Omega\times\Omega}h(x-y)d\rho(x)d\rho(y)$ 为 W_2 凸的。

证明: 对于任意 $\mathbb{W}_2$ 空间上的恒速测地线 $\rho:[0,1]\to\mathscr{P}_2(\Omega)$, 总存在 $\rho_0\to\rho_1$ 的最优传输方案 γ，使得 $\rho_t=(\pi_t)\sharp\gamma,\pi_t(x,y)=(1-t)x+ty$. 此时：

$$\int_\Omega V d\rho_t = \int_\Omega V((1-t)x+ty)d\gamma(x,y)$$

$$\iint_{\Omega\times\Omega} h(x-y)d\rho_t(x)d\rho_t(y)$$
$$= \iint_{\Omega\times\Omega} h((1-t)(x_1-y_1)+t(x_2-y_2))d\gamma(x_1,x_2)d\gamma(y_1,y_2)$$

而若 V 为凸时，$t\mapsto V((1-t)x+ty)$ 将为凸函数，因此 $t\mapsto \mathcal{F}[\rho_t]=\int_\Omega V((1-t)x+ty)d\gamma(x,y)$ 为凸函数。同理，若 h 为凸时，$t\mapsto h((1-t)(x_1-y_1)+t(x_2-y_2))$ 为凸函数，因此 $t\mapsto \mathcal{F}[\rho_t]=\iint_{\Omega\times\Omega} h(x-y)d\rho_t(x)d\rho_t(y)$ 为凸，即得上述两个泛函 $\mathcal{F}[\rho]$ 为 W_2 凸的。 ■

9 Gradient Flow

简单的手写笔记

本章参考的资料为:An-introduction-to-optimal-transport-and-Wasserstein-gradient-flows(Figalli)，及其扩展版本An Invitation to Optimal Transport, Wasserstein Distances, and Gradient Flows

9.1 欧式空间下的梯度流

定义（梯度流）：设 $\mathscr{H}$ 为 Hilbert 空间，$\phi:\mathscr{H}\to\mathbb{R}\in C^1$, 即 ϕ 连续可微。则关联 ϕ，起点为 x_0 的梯度流为如下常微分方程（ODE）的解

$$\begin{cases} x(0)=x_0, \\ \dot{x}(t)=-\nabla\phi(x(t)) \end{cases}$$

由上述定义知，若曲线 $x(t), t\in[0,T)$ 为 ϕ 的梯度流，则:

$$\frac{d}{dt}\phi(x(t))=\nabla\phi(x(t))\cdot\dot{x}(t)=-|\nabla\phi|^2(x(t))\le 0$$

因此，ϕ 沿着曲线 $x(t)$ 递减；$\frac{d}{dt}\phi(x(t))=0$ 当且仅当 $\nabla\phi(x(t))=0$, 也就是 $x(t)$ 为 $\phi(x)$ 的邻接点。若 ϕ 为强凸函数（有唯一的全局极小点，且为临界点），则 $t\to\infty, x(t)$ 将收敛与此。

求解上述 ODE，换言之，构建梯度流，可采用隐式欧拉（反向欧拉）方法。给定某个时刻 t, 以及 $x(t)$, 固定一步长 $\tau>0$, 则寻找一个 $x(t+\tau)\in\mathscr{H}$ 满足:

$$\frac{x(t+\tau)-x(t)}{\tau}=-\nabla\phi(x(t+\tau))$$

基于上述思想，设置 $x_0^\tau=x_0$, 给定 $k\ge 0$ 与 x_k^τ, 则 x_{k+1}^τ 通过求解如下方程获得:

$$\frac{x_{k+1}^\tau-x_k^\tau}{\tau}=-\nabla\phi(x_{k+1}^\tau)$$

可等价的写成如下形式：

$$\nabla_x\left(\frac{\|x-x_k^\tau\|^2}{2\tau}+\phi(x)\right)\Big|_{x=x_{k+1}^\tau}=\frac{x_{k+1}^\tau-x_k^\tau}{\tau}+\nabla\phi(x_{k+1}^\tau)=0$$

因此把求解梯度流（离散化后的每个时间步）等价于求解优化问题

$$x_{k+1}^\tau=\arg\min_x(\frac{\|x-x_k^\tau\|^2}{2\tau}+\phi(x))$$

这正是求解 $\min_x\phi(x)$ 的 Proximal Point Method 的每个迭代步。

Note: 上式取代 $\|\cdot\|^2/2$ 以 Bregman 散度变为镜像下降，若取代正则项以 $\mathcal{W}_2^2(\cdot,\cdot)$，则将欧式空间梯度流变为 Wasserstein 梯度流。

可定义 ϕ 的微分（方向导数）如下，

$$d\phi(x)[v]=\lim_{\varepsilon\to0}\frac{\phi(x+\varepsilon v)-\phi(x)}{\varepsilon}$$

由于：

$$\dot{x}(t)=\lim_{\varepsilon\to0}\frac{x(t+\varepsilon)-x(t)}{\varepsilon}\in\mathcal{H}$$

则可定义标量积 $<\cdot,\cdot>:\mathscr{H}\times\mathscr{H}\to\mathscr{H}^*$ 如下（其中 $\mathscr{H}^*$ 为 $\mathscr{H}$ 的（Legendre）对偶空间）：

$$\langle\nabla\phi(x),v\rangle:=d\phi(x)[v]\quad\forall v\in\mathcal{H}$$

这也可以看作 Sec.8.2, 构建 $\mathbb{W}_2$ 空间下相对应的标量积（内积）、梯度以构建 Wasserstein 梯度流的出发点。

上述定义假设 $\phi\in C^1$, 实际问题没有这么理想，因此梯度流天然有一个推广定义：

定义（梯度流 2）： 设 $\mathscr{H}$ 为 Hilbert 空间，$\phi:\mathscr{H}\to\mathbb{R}$ 为下半连续函数 ($\liminf_{y\to x}f(y)\geq f(x)$，减去一个小正数 ϵ 后小于其一小邻域内的任一点函数值)。称一绝对连续曲线（一条几乎处处可微的连续曲线）$x:[0,+\infty]\to\mathscr{H}$ 为关联 ϕ，起点为 x_0 的梯度流，如果满足如下常微分方程

$$\begin{cases}x(0)=x_0,\\ \dot{x}(t)=-\partial\phi(x(t))\end{cases}$$

其中 $\partial\phi(x(t))$ 为次梯度。

类似的，采用隐式 Euler 构建上述梯度流。置 $x_0^\tau=x_0$，给定 $k\geq0$ 与 x_k^τ，则 x_{k+1}^τ 通过求解如下方程获得：

$$\frac{x_{k+1}^\tau-x_k^\tau}{\tau}\in-\partial\phi(x_{k+1}^\tau)$$

则其等价形式为：

$$0 \in \frac{x_{k+1}^\tau - x_k^\tau}{\tau} + \partial\phi(x_{k+1}^\tau) =: \partial\psi_k^\tau(x_{k+1}^\tau), \quad \psi_k^\tau(x) := \frac{\|x - x_k^\tau\|^2}{2\tau} + \phi(x)$$

有了梯度流的定义，我们现在讨论以下其稳定性和唯一性。

我们有如下结论，若 ϕ 为凸函数，$x(t), y(t)$ 分别为以 x_0, y_0 为起点的关于 ϕ 的梯度流，则

1. 若 $x_0 = y_0$，则 $x(t) = y(t)$，即唯一性

2. 若 $x_0 \to y_0$，则 $\forall t \in [0, +\infty]$,$x(t), y(t)$ 保持一致接近，即稳定性

上述稳定性与唯一性完全依赖于 ϕ 的凸性。以 $\phi \in C^1$ 这种简单情况来说明，

$$\begin{aligned}\frac{d}{dt}\frac{\|x(t) - y(t)\|^2}{2} &= \langle x(t) - y(t), \dot{x}(t) - \dot{y}(t)\rangle \\ &= -\langle x(t) - y(t), \nabla\phi(x(t)) - \nabla\phi(y(t))\rangle \le 0\end{aligned}$$

即 $x(t)$ 与 $y(t)$ 之间的偏差随时间演化递减。

最后，我们给一个例子：Dirichlet 能量泛函的 L^2 梯度流，其正为传热方程（Heat Equation）。

Note: 我们将在下一节说明传热方程的另一等价形式，即熵泛函的 Wasserstein 梯度流。另外 Dirichlet 能量泛函的 Euler-Lagrange 方程为 Laplace 方程，与传热方程都含有 Δu, 这一部分请参考附录A的手写笔记。

命题： 设 $\mathscr{H} = L^2(\mathbb{R}^2)$, 且

$$\phi(u) = \begin{cases} \frac{1}{2}\int_{\mathbb{R}^d} |\nabla u|^2 dx & \text{if} u \in W^{1,2}(\mathbb{R}^d), \\ +\infty & \text{otherwise.} \end{cases}$$

则：$\partial\phi(u) \neq \emptyset \quad \Leftrightarrow \quad \Delta u \in L^2(\mathbb{R}^d)$. 且 $\partial\phi(u) = \{-\Delta u\}$.

证明： "⇒". 由于 $\partial\phi(u) \neq \emptyset$, 取 $p \in L^2(\mathbb{R}^2), p \in \partial\phi(u)$. 由次微分的定义有：

$$\phi(v) \ge \phi(u) + \langle p, v - u\rangle_{L^2}, \forall v \in \mathscr{H}$$

取 $v = u + \varepsilon w, w \in W^{1,2}(\mathbb{R}^d), \varepsilon > 0$. 带入上式得：

$$\int_{\mathbb{R}^d} \frac{|\nabla(u + \varepsilon w)|^2}{2} dx - \int_{\mathbb{R}^d} \frac{|\nabla u|^2}{2} dx \ge \varepsilon \int_{\mathbb{R}^d} pw dx$$

化简得:

$$\int_{\mathbb{R}^d} \nabla u \cdot \nabla w dx + \frac{\varepsilon}{2}\int_{\mathbb{R}^d} |\nabla w|^2 dx \ge \int_{\mathbb{R}^d} pw dx$$

对 ε 取极限，有保号性得：

$$\int_{\mathbb{R}^d} \nabla u \cdot \nabla w \geq \int_{\mathbb{R}^d} pwdx \quad \forall w \in W^{1,2}(\mathbb{R}^d)$$

取 $w = -w$, 因此不等号中将变为等号，且对左侧分部积分（或 Stokes 公式），有：

$$\int_{\mathbb{R}^d} \underbrace{-\Delta u}_{\substack{\text{as a} \\ \text{distribution}}} wdx = \int_{\mathbb{R}^d} \nabla u \cdot \nabla w dx = \int_{\mathbb{R}^d} pwdx \quad \forall w \in W^{1,2}(\mathbb{R}^d)$$

由 $w \in W^{1,2}(\mathbb{R}^d)$ 的任意性知：$-\Delta u = p \in L^2(\mathbb{R}^2)$. 且 $p \in \partial\phi(u)$ 的任意性知，$\partial\phi(u) = \{-\Delta u\}$

"$\Leftarrow$", 即证 $-\Delta u \in \partial\phi(u)$ 即可。设 $-\Delta u \in L^2(\mathbb{R}^2)$,$\forall w \in W^{1,2}(\mathbb{R}^d)$, 有：

$$\begin{aligned}\phi(u+w) - \phi(u) &= \int_{\mathbb{R}^d} \nabla u \cdot \nabla w dx + \frac{1}{2}\int_{\mathbb{R}^d} |\nabla w|^2 dx \\ &\geq \int_{\mathbb{R}^d} \nabla u \cdot \nabla w dx = \int_{\mathbb{R}^d} -\Delta u w dx =< -\Delta u, w >\end{aligned}$$

另一方面，当 $w \notin W^{1,2}(\mathbb{R}^d)$ 时，

$$\phi(u+w) = +\infty \geq \phi(u) + \int_{\mathbb{R}^d} -\Delta u w dx =< -\Delta u, w >$$

由次微分的定义知 $-\Delta u \in \partial\phi(u)$.

有了上述命题，再加上 ϕ 的梯度流满足的微分方程为：$\partial_t u(t,x) \in -\partial\phi(u)$, 而 $\partial\phi(u) = \{-\Delta u\}$ 即得 $\partial_t u(t,x) = \Delta u.$, 即如下推论

推论（传热方程为梯度流）： 设 $\mathscr{H} = L^2(\mathbb{R}^2)$, 考虑 Dirichlet 能量泛函

$$\phi(u) := \begin{cases} \frac{1}{2}\int_{\mathbb{R}^d} |\nabla u|^2 dx & if u \in W^{1,2}(\mathbb{R}^d), \\ \\ +\infty & otherwise. \end{cases}$$

则：

$$\partial_t u(t) \in -\partial\phi(u(t)) \quad \Leftrightarrow \quad \partial_t u(t,x) = \Delta u(t,x)$$

即 Dirichlet 能量泛函 $\phi(u)$ 相对于 L^2 标量积的梯度流等价于传热方程。

因此，构建传热方程的解，可以通过构建 Dirichlet 能量泛函的梯度流的方式获得。而构建梯度流，则采用上述隐式 Euler 法，转化为求一系列优化问题：

$$u_{k+1}^\tau \text{ is the minimizer in } L^2(\mathbb{R}^d) \text{ of } u \mapsto \frac{\|u - u_k^\tau\|_{L^2(\mathbb{R}^d)}^2}{2\tau} + \phi(u), \text{且要求} \tau \to 0$$

这种将构建 PDE 解的方式，作用于下一节构建 Wasserstein 梯度流情况，求解一系列优化问题：

$$\rho_{(k+1)}^\tau \in \text{argmin}_\rho \mathcal{F}(\rho) + \frac{W_2^2(\rho, \rho_{(k)}^\tau)}{2\tau}$$

称为 JKO 框架.

9.2 Wasserstein Gradient Flow 与 JKO Scheme

本小节的内容主要围绕如下两句话展开:

Fokker Plank Equation of the Langevin Diffusion = Wasserstein Gradient Flow Associated with the KL Divergence.

Heat Equation is the Gradient Flow of the Entropy Functional with respect to the W2 Metric.

定义（Wasserstein Gradient Flow）: 给定泛函 $\mathcal{F}:\mathscr{P}_2(\Omega)\to\mathbb{R}$, 称概率测度曲线 $\rho:[0,T)\to\mathscr{P}_2(\Omega)$ 为泛函 $\mathcal{F}$ 相对于 W_2 距离以 $\bar{\rho}_0$ 为起点的梯度流，如果满足如下方程 (简记 $\rho(t)=\rho_t$ 为以概率密度（或概率测度）):

$$\begin{cases}\partial_t\rho_t=-\mathrm{grad}_{W_2}\mathcal{F}[\rho_t]=\mathrm{div}\Big(\nabla\Big(\frac{\delta\mathcal{F}[\rho_t]}{\delta\rho_t}\Big)\rho_t\Big)\\ \rho_0=\bar{\rho}_0\end{cases}$$

现在，我们从最优传输角度求解如下一系列优化问题:

$$\rho^{\tau}_{(k+1)}\in\mathrm{argmin}_{\rho}\mathcal{F}(\rho)+\frac{W_2^2(\rho,\rho^{\tau}_{(k)})}{2\tau}$$

推导/构建出上述 wasserstein 梯度流。

根据右侧取最小，则（参考文献 [1]）

$$\frac{\delta F}{\delta\rho}(\rho)+\frac{\phi}{\tau}=constant$$

其中:ϕ 为成本函数为 $|x-y|/2$ 时的 Kantorovich 势能函数。上式对 x 求偏导: $-\frac{\nabla\phi(x)}{\tau}=\nabla\big(\frac{\delta\mathcal{F}}{\delta\rho}(\rho)\big)(x)$. 根据 Brenier 定理，$T(x)=x-\nabla\phi(x)$，因此

$$\frac{T(x)-x}{\tau}=\nabla\big(\frac{\delta\mathcal{F}}{\delta\rho}(\rho)\big)(x)$$

设速度 $-\mathbf{v}=\frac{T(x)-x}{\tau}$, 有连续方程 $\partial_t\rho_t+\nabla\cdot(\rho_t v_t)=0$

带入 $\mathbf{v}$ 即得: $\partial_t\rho_t=\mathrm{div}\Big(\nabla(\frac{\delta\mathcal{F}[\rho_t]}{\delta\rho_t})\rho_t\Big)$

- 取 $\mathcal{F}[\rho]=\int\rho\log\rho dx$, 有: $\mathrm{grad}_{W_2}\mathcal{F}[\rho]=-\Delta\rho$ 则 Wasserstein Gradient Flow 变为 Heat Equation:

$$\begin{cases}\partial_t\rho_t=\Delta\rho_t\\ \rho_0=\bar{\rho}_0\end{cases}$$

- 取 $\mathcal{F}[\rho]=\int_{\mathbb{R}^d}\rho\log(\rho)+\rho V(x)dx$。当然，可以做一些变量代换，

$$\mathcal{F}[\rho]=\int_{\mathbb{R}^d}\eta\log(\eta)\cdot e^{-V}dx,\quad \eta:=e^V\rho$$

此时，可以看出 $\mathcal{F}[\rho] \geq 0$, 这是因为

$$\begin{aligned}\mathcal{F}[\rho] = \int_{\mathbb{R}^d} \eta \log(\eta) e^{-V} dx &\geq \left(\int_{\mathbb{R}^d} \eta e^{-V} dx\right) \log\left(\int_{\mathbb{R}^d} \eta e^{-V} dx\right) \\ &= \left(\int_{\mathbb{R}^d} \rho dx\right) \log\left(\int_{\mathbb{R}^d} \rho dx\right) = 0\end{aligned}$$

上述不等式为凸函数的 Jessen 不等式：$\mathbb{E}f(x) \geq f(\mathbb{E}(x))$。当 $\eta = 1$ 时，$\mathcal{F}[\rho]$ 达到最小，对应的 $\rho^* = e^{-V}$。

回到正题，此时 $\mathrm{grad}_{W_2}\mathcal{F}[\rho] = -\Delta\rho - \mathrm{div}(\rho\nabla V) = -\mathrm{div}(\nabla\rho + \rho\nabla V)$，则其 Wasserstein norm 为：

$$\begin{aligned}\langle \mathrm{grad}_{W_2}\mathcal{F}[\rho], \mathrm{grad}_{W_2}\mathcal{F}[\rho]\rangle_\rho &= \int_{\mathbb{R}^d} \left|\frac{\nabla\rho + \rho\nabla V}{\rho}\right|^2 \rho dx \\ &= \int_{\mathbb{R}^d} \frac{\left|e^{-V}\nabla\eta - \eta\nabla V e^{-V} + \eta e^{-V}\nabla V\right|^2}{e^{-V}\eta} dx \\ &= \int_{\mathbb{R}^d} \frac{|\nabla\eta|^2}{\eta} e^{-V} dx\end{aligned}$$

关联的 Wasserstein Gradient Flow 变为 Fokker–Planck Equation:

$$\begin{cases}\partial_t\rho_t = \Delta\rho_t + \mathrm{div}(\nabla V\rho_t) \\ \rho_0 = \bar{\rho}_0\end{cases}$$

这正好也是过阻尼 Langevin 方程 $dX = -\nabla V(X)dt + \sqrt{2}dW$, 其中 W 为 Brownian Motion, 的 Fokker–Planck Equation（又称 SDE 的 Kolmogorov Forward Equation）. 其物理含义为：描述了 Brownian Motion 在存在外部保守力下的速度演化，注意 X 在这里不是位置而是速度。更一般的 Langevin 方程还可以加上阻尼项 $dX = -bX - \nabla V(X)dt + \sqrt{2}dW$,b 为（摩擦）阻尼系数.

特别地，取 $V = -\log q(x)$,q 为另一概率密度，则 $\mathcal{F}[\rho] = \int_{\mathbb{R}^d} \rho\log\frac{\rho}{q}dx = KL(\rho||q)$, 则 Wasserstein Gradient Flow 变为：

$$\begin{cases}\partial_t\rho_t = \Delta\rho_t - \mathrm{div}(\rho_t\nabla\log q(x)) \\ \rho_0 = \bar{\rho}_0\end{cases}$$

对应的 Langevin 方程为 $dX = \nabla\log q(x)dt + \sqrt{2}dW$. 这正是深度学习中 Diffusion Model(一种 Score Based Model) 中的朗之万采样方程, 其中称 $\nabla\log q(x)$ 为 Score。采用 JKO 框架，通过求解

$$\rho^\tau_{(k+1)} \in \mathrm{argmin}_\rho \mathcal{F}(\rho) + \frac{W_2^2(\rho, \rho^\tau_{(k)})}{2\tau}$$

, 构建梯度流。可以看到本质上就在采用 Proximal Point Method 求解 $\arg\min_\rho \mathcal{F}(\rho)$ 问题，而此问题的最优解为 $\rho^*(x) = q(x)$。因此，此梯度流将收敛于概率密度 $q(x)$，这就是通过朗之万采样，可以获取给定分布 $q(x)$ 的样本的原因。抑或直接将 Langevin 方程视为求解微分方程的 Euler 法，x 将在梯度 $\nabla \log q(x)$ 的作用下，沿 $q(x)$ 增大的方向移动，并因有 dW 项而有随机性，因而能很好的采样出 $q(x)$ 分布下的样本。

上述通过求解一系列优化问题：

$$\rho^\tau_{(k+1)} \in \text{argmin}_\rho \mathcal{F}(\rho) + \frac{W_2^2(\rho, \rho^\tau_{(k)})}{2\tau}$$

，换言之通过构建某个泛函的 Wasserstein Gradient Flow，来构造一些偏微分方程 (PDE)，如传热方程、Fokker-Planck 方程，解的框架称为 **Jordan-Kinderlehrer-Otto scheme**(JKO scheme,1998).

9.2.1 Entropy Functional 与 Heat Equation

在本节的绪论部分已经把结论给解释清楚了：传热方程等价于熵泛函的 Wasserstein 梯度流。我们不加证明（见 [9] Page 75）的给出

定理： 给定 $\tau > 0$, 设 $\rho^\tau : [0, +\infty) \to \mathscr{P}_2(\Omega)$ 为概率测度曲线, 定义如下：

$$\rho^\tau(t) := \begin{cases} \bar{\rho}_0 \quad for \quad t = 0, \\ \rho^\tau_k \quad for \quad t \in ((k-1)\tau, k\tau], k \geq 1 \end{cases}$$

其中：

$$\rho^\tau_{k+1} \in \text{argmin}_\rho \int_\Omega \rho \log(\rho) dx + \frac{W_2^2(\rho, \rho^\tau_k)}{2\tau}$$

则存在一绝对连续概率测度曲线 $\rho \in L^1_{\text{loc}}([0, \infty) \times \Omega)$, 使得 $\rho^\tau \rightharpoonup \rho$(弱收敛)。并且，$\rho$ 满足以初始条件为 ρ_0，边界为 *zero Neumann* 边界条件的传热方程（在分布意义下）：$\partial_t \rho(t, x) = \Delta \rho(t, x)$.

最后，不加证明的给出文献 [9]Page 73 页关于梯度流构建每一步所求的最优 ρ^τ_{k+1} 所满足的性质。

引理： 对于任意与边界 $\partial\Omega$ 相切的向量场 $\xi \in C^\infty(\Omega, \mathbb{R}^n)$，$\rho^\tau_{k+1}$ 满足：

$$\int_\Omega \rho^\tau_{k+1} \,\text{div}(\xi) dx = \frac{1}{\tau} \int_\Omega \langle \xi \circ T_{k+1}, T_{k+1} - x \rangle \rho^\tau_k dx$$

其中 $\rho^\tau_{k+1} \in \text{argmin}_\rho \int_\Omega \rho \log(\rho) dx + \frac{W_2^2(\rho, \rho^\tau_k)}{2\tau}$,$T_{k+1} : \Omega \to \Omega$ 为 $\rho^\tau_k \to \rho^\tau_{k+1}$ 的最优传输映射。

9.2.2 KL 散度、Fokker-Planck Equation 与 Langevin Equation

我们根据文献 [9] 的 4.4 节给出泛函 $\mathcal{F}[\rho] = \int_{\mathbb{R}^d} \rho\log(\rho) + \rho V(x)dx$ 为 $\lambda-$ 凸情况下, 对应的 Fokker-Planck Equation 解的收敛性与 *contractivity*。

我们先给出之前的结论: 当取泛函为: $\mathcal{F}[\rho] = \int_{\mathbb{R}^d} \rho\log(\rho) + \rho V(x)dx$ 时, 其 Wasserstein 梯度流为如下 Fokker-Planck 方程:

$$\begin{cases} \partial_t\rho_t = -\mathrm{grad}_{W_2}\mathcal{F}[\rho_t] = \mathrm{div}\Big(\nabla\Big(\frac{\delta\mathcal{F}[\rho_t]}{\delta\rho_t}\Big)\rho_t\Big) = \Delta\rho_t + \mathrm{div}(\nabla V\rho_t) \\ \rho_0 = \bar{\rho}_0 \end{cases}$$

由于

$$\frac{d}{dt}\mathcal{F}[\rho_t] = \langle \mathrm{grad}_{W_2}\mathcal{F}[\rho_t], \partial_t\rho_t\rangle_{W_2} = -\langle \mathrm{grad}_{W_2}\mathcal{F}[\rho_t], \mathrm{grad}_{W_2}\mathcal{F}[\rho_t]\rangle_{\rho_t} \le 0$$

因此，随着梯度流的演化，$\mathcal{F}[\rho]$ 将逐渐下降。由于 $\min\mathcal{F}[\rho] = \mathcal{F}[e^{-V}]$. 那么如果 $\mathcal{F}[\rho]$ 有强凸的性质，则

$$\mathcal{F}[\rho_t] \to \mathcal{F}[e^{-V}] = 0 \quad \text{as} \quad t \to \infty$$

我们这里的强凸采用 $\lambda-$ 凸刻画。

之前也提到过 $\phi : \mathbb{R}^n \to \mathbb{R}$ 为 $\lambda-$ 凸的含义，即 $\phi - \lambda\frac{\|x\|^2}{2}$ 仍为凸函数。在此我们将其适应到测度空间 $\mathscr{P}_2(\Omega)$ 上。

我们用凸函数的一阶定义重新描述 $\lambda-$ 凸函数的定义

定义（$\lambda-$ 凸函数）: 设 $I \subset \mathbb{R}$, 下半连续函数 $\phi : I \to \mathbb{R}\cup\{+\infty\}$, 给定 $\lambda > 0$, 称 ϕ 为 $\lambda-$ 凸函数如果:

$$\begin{aligned}&(1-s)\varphi(x) + s\varphi(y) \\ &\qquad \ge \varphi((1-s)x + sy) + \frac{\lambda s(1-s)}{2}|x-y|^2 \quad \forall x, y \in I, 0 \le s \le 1\end{aligned}$$

在一般的测地度量空间 (X, d)，如 $(\mathscr{P}(\Omega), W_2)$ 上，称其上的下半连续函数 $\phi : X \to \mathbb{R}\cup\{+\infty\}$ 为 $\lambda-$ 凸, 当且仅当给定 X 上的任意测地线 $\gamma : [0,1] \to X$, $\phi\circ\gamma$:$[0,1] \to \mathbb{R}\cup\{+\infty\}$ 均为 $\lambda-$ 凸。

上述条件和 $\phi - \lambda\frac{\|x\|^2}{2}$ 为凸函数一致。因此，假设 ϕ 是 λ-凸，则利用 $\phi - \lambda\frac{\|x\|^2}{2}$ 为凸函数，以及凸函数的定义有:

$$\phi(y) \ge \phi(x) + \langle\nabla\phi(x), y-x\rangle + \frac{\lambda}{2}|y-x|^2 \quad \forall x, y \in \mathbb{R}^d$$

交换 x, y 位置, 可以进一步获得:

$$\langle x-y, \nabla\phi(x) - \nabla\phi(y)\rangle \ge \lambda|x-y|^2$$

引理（$\lambda-$ 凸函数的可加性）： 若 $\phi_i, i=1,2$ 分别为 λ_1- 凸函数,λ_2- 凸函数，则 $\phi_1+\phi_2$ 为 $(\lambda_1+\lambda_2)-$ 凸函数.

考虑 $\lambda-$ 凸函数 $\phi(x)\in C^1(\mathbb{R}^n,\mathbb{R})$, 设其最小点为 x^*, 有 $\nabla\phi(x^*)=0$, 则有：

$$\phi(x)\geq\phi(x^*)+\frac{\lambda}{2}|x-x^*|^2\Rightarrow\sqrt{\frac{2}{\lambda}\big(\phi(x)-\phi(x^*)\big)}\geq|x-x^*|$$

应用 ϕ 的凸性，有：

$$\begin{aligned}\phi(x^*)&\geq\phi(x)+\langle\nabla\phi(x),x^*-x\rangle+\frac{\lambda}{2}|x-x^*|^2\\&\Rightarrow\langle\nabla\phi(x),x-x^*\rangle\geq\phi(x)-\phi(x^*)+\frac{\lambda}{2}|x-x^*|^2\end{aligned}$$

因此：$|\nabla\phi(x)|\geq\frac{\phi(x)-\phi(x^*)}{|x-x^*|}+\frac{\lambda}{2}|x-x^*|\geq\sqrt{2\lambda\left(\phi(x)-\phi(x^*)\right)}$

把上述结论推广到泛函 $\mathcal{F}[\rho]$ 上有：

引理： 给定一下半连续且 $\lambda-$ 凸的泛函 $\mathcal{F}[\rho]:\mathscr{P}_2(\mathbb{R}^n)\to\mathbb{R}\cup\{+\infty\}$，$\lambda>0$. 设 $\min\mathcal{F}=\mathcal{F}[\bar{\rho}]$, 则对 $\forall\rho\in\mathscr{P}_2(\mathbb{R}^n)$, 有如下两式成立：

$$\begin{aligned}W_2^2(\rho,\bar{\rho})&\leq\frac{2}{\lambda}\big(\mathcal{F}[\rho]-\mathcal{F}[\bar{\rho}]\big),\\\mathcal{F}[\rho]-\mathcal{F}[\bar{\rho}]&\leq\frac{1}{2\lambda}\langle\mathrm{grad}_{W_2}\mathcal{F}[\rho],\mathrm{grad}_{W_2}\mathcal{F}[\rho]\rangle_\rho\end{aligned}$$

命题： 设 $V:\mathbb{R}^n\to\mathbb{R}$ 为 $\lambda-$ 凸函数，则泛函 $\mathcal{F}[\rho]=\int_{\mathbb{R}^d}\rho\log(\rho)+\rho V dx$ 为 $\lambda-$ 凸函数.

现在，我们取 $\lambda-$ 凸的 V,$\lambda>0$, 证明收敛性：

$$\mathcal{F}[\rho_t]\to\mathcal{F}[e^{-V}]=0\quad\text{as}\quad t\to\infty$$

结合上述结果，有

$$\frac{d}{dt}\mathcal{F}[\rho_t]=-\langle\mathrm{grad}_{W_2}\mathcal{F}[\rho_t],\mathrm{grad}_{W_2}\mathcal{F}[\rho_t]\rangle_{\rho_t}\leq-2\lambda(\mathcal{F}[\rho_t]-\mathcal{F}[e^{-V}])=-2\lambda\mathcal{F}[\rho_t]$$

因此有 $\frac{d}{dt}\big(\mathcal{F}[\rho_t]e^{2\lambda t}\big)=\frac{d}{dt}\mathcal{F}[\rho_t]\cdot e^{2\lambda t}+2\lambda e^{2\lambda t}\mathcal{F}[\rho_t]\leq0$

因此 $\mathcal{F}[\rho_t]$ 将至少以 $e^{-2\lambda t}$ 这一指数级进行进行衰减！因此 $\mathcal{F}[\rho_t]\leq e^{-2\lambda t}\mathcal{F}[\bar{\rho}_0]$, 随 $t\to+\infty$, 最终必收敛于 $\mathcal{F}[e^{-V}]=0$.

对 ρ_t 用一下上述引理我们可以获得：

$$W_2^2(\rho_t,e^{-V})\leq\frac{2}{\lambda}\mathcal{F}[\rho_t]\leq\frac{2}{\lambda}\mathcal{F}[\bar{\rho}_0]e^{-2\lambda t}\quad\forall t\geq0$$

再对 $\bar{\rho}_0$ 用一下引理结果有：$W_2^2(\bar{\rho}_0,e^{-V})\leq\frac{2}{\lambda}\mathcal{F}[\bar{\rho}_0]$.

我们用 $W_2^2(\bar{\rho}_0, e^{-V})$ 取代 $\frac{2}{\lambda}\mathcal{F}[\bar{\rho}_0]$ 将获得一个更强的版本：

$$W_2^2(\rho_t, e^{-V}) \leq e^{-2\lambda t} W_2^2(\bar{\rho}_0, e^{-V})$$

此更强版本是由 $\mathcal{F}[\rho]$ 的凸性保证的（不知道咋推出来的）.

此公式还能进行推广，用关联于泛函 $\mathcal{F}[\rho]$ 一新的梯度流（也就是一条新的概率测度曲线）$\tilde{\rho}:[0,+\infty)\to\mathscr{P}_2(\mathbb{R}^n)$ 去近似替换 e^{-V}, 有如下公式成立：

$$W_2^2(\rho_t, \tilde{\rho}_t) \leq e^{-2\lambda t} W_2^2(\rho_0, \tilde{\rho}_0) \quad \forall t \geq 0$$

这表明：若同一 $\lambda-$ 凸泛函关联的两梯度流，若初始值相近，则两梯度流将一致接近（也就是稳定性），若初值相同，则两梯度流唯一（唯一性），与上一小节欧氏空间 $\mathbb{R}^n$ 的结果一致。因此，这一属性又称为梯度流的 *contractivity* (收缩性)。

关于其证明，参考文献 [10]。下面简单用 $\mathbb{R}^n$ 下的梯度流进行推导。

设 $\varphi(x)\ \mathbb{R}^n\to\mathbb{R}$ 为 $\lambda-$ 凸函数, 考虑其两条梯度流 $x(t), y(t)$, 其满足

$$\dot{x}(t) = -\nabla\varphi(x(t)), \quad \dot{y}(t) = -\nabla\varphi(y(t))$$

则

$$\begin{aligned}\frac{d}{dt}\frac{|x(t)-y(t)|^2}{2} &= \langle x(t)-y(t), \dot{x}(t)-\dot{y}(t)\rangle \\ &= -\langle x(t)-y(t), \nabla\varphi(x(t))-\nabla\varphi(y(t))\rangle \\ &\leq -\lambda|x(t)-y(t)|^2\end{aligned}$$

最后一个不等式，利用了 φ 的 $\lambda-$ 凸性：

$$\varphi(y) \geq \varphi(x) + \langle\nabla\varphi(x), y-x\rangle + \frac{\lambda}{2}|y-x|^2$$

再交换一下 x, y，然后再加起来得：$\langle x-y, \nabla\varphi(x)-\nabla\varphi(y)\rangle \geq \lambda|x-y|^2$.

最终可得：

$$\frac{d}{dt}\left(|x(t)-y(t)|^2 e^{2\lambda t}\right) \leq 0 \Rightarrow |x(t)-y(t)|^2 \leq e^{-2\lambda t}|x(0)-y(0)|^2$$

Remark（Langevin 方程的收敛性）：由于 Langevin 方程:

$$\frac{dX_t}{dt} = -\nabla V(X_t) + \sqrt{2}\frac{dB_t}{dt}$$

的 Kolmogorov 前向方程就是泛函 $\mathcal{F}[\rho] = \int_{\mathbb{R}^d}\rho\log(\rho) + \rho V(x)dx$ 的 Wasserstein 梯度流方程。因此上述的结果也在 Langevin 方程所描述的随机过程（或者采样路径）上成立。

记：$\frac{dX_t}{dt} = -\nabla V(X_t) + \sqrt{2}\frac{dB_t}{dt}$，$\frac{dY_t}{dt} = -\nabla V(Y_t) + \sqrt{2}\frac{dB_t}{dt}$，$\alpha_t = X_t - Y_t$, 则：$\frac{d\alpha_t}{dt} = -(\nabla V(X_t) - \nabla V(Y_t))$. 因此

$$\frac{d}{dt}\frac{|\alpha_t|^2}{2} = -\langle \nabla V(X_t) - \nabla V(Y_t), X_t - Y_t \rangle \leq -\lambda |X_t - Y_t|^2 = -\lambda |\alpha_t|^2$$

不等号利用了上述 V 的 $\lambda-$ 凸性。再根据 Gronwall 不等式的微分形式，有

$$|\alpha_t|^2 \leq e^{-2\lambda t}|\alpha_0|^2$$

再进一步假设 $\mathbb{E}[X_0^2]$,$\mathbb{E}[Y_0^2]$ 有限，则

$$\mathbb{E}|X_t - Y_t|^2 \leq e^{-2\lambda t}\mathbb{E}|X_0 - Y_0|^2 \leq 2\left(\mathbb{E}|X_0|^2 + \mathbb{E}|Y_0|^2\right)e^{-2\lambda t}$$

这说明 $\alpha_t = X_t - Y_t, t \to +\infty$ 最终将收敛至 0.

Gronwall's Inequality:

- 设 ϕ, f $[o, T] \to \mathbb{R}$, 非负、连续，C_0 为常数。若

$$\phi(t) \leq C_0 + \int_0^t f\phi ds \quad \forall 0 \leq t \leq T$$

则：

$$\phi(t) \leq C_0 e^{\int_0^t f ds} \quad \forall 0 \leq t \leq T$$

- 设 $\phi, \psi, \eta : [o, T] \to \mathbb{R}$ 非负，η 绝对连续，且 a.e. t, 满足：

$$\eta'(t) \leq \phi(t)\eta(t) + \psi(t)$$

则：

$$\eta(t) \leq e^{\int_0^t \phi(s)ds}\left[\eta(0) + \int_0^t \psi(s)ds\right] \quad \forall 0 \leq t \leq T$$

9.3 Numerical Method

9.3.1 Entropic wasserstein gradient flows

先介绍一个完全基于“传统”优化的方法Entropic wasserstein gradient flows, 然后介绍基于深度学习的方法。

9.3.2 JOKNet:Learning the Energy Functional

本节参考Proximal Optimal Transport Modeling of Population Dynamics(JKOnet)

9.3.3 Computing Wasserstein Gradient Flows with ICNN

Large-Scale Wasserstein Gradient Flows

9.3.4 Variational Wasserstein Gradient Flow

参考Variational Wasserstein gradient flow, 作者Jiaojiao Fan. 之前已介绍她利用 Input Convex Neural Networks 求解 Wasserstein Barycenter 的工作。他的导师为Yongxin Chen, 这两位甚至都是机动校友。。。机动还有这种人才。。。

10 Mean Field Games

极简单手写笔记 MFG,一个与手写笔记形式相同的笔记。更一般的资料：Note on Mean Field Games, notes-mean-field, Variational Mean Field Games and Optimal Transport. 以及上述笔记的“合订本” An Introduction to Mean Field Game Theory; 当然还有其他的书，如Probabilistic Theory of Mean Field Games with ApplicationsI、II

10.1 概述与动机

想象一个场景：在酷暑的夏天，你怀着不甘窝在家里的心态心血来潮的独自去往了海滩避暑。到了中午，天太热了，你你想找一个阴凉的地方避暑，当然沙滩上不长枝繁叶茂的树，但是有一个较大的遮阳伞，然而沙滩上有和你相同很多。我们想要避暑，因此我们会选择到遮阳伞下面，然而伞大小有限，人太多了反而更热，因此你不得不离开遮阳伞，如果从遮阳伞走的人多了，那相比于在外面晒着，自然还是回到遮阳伞下面更好。

如上的场景是一个多人的博弈 (*games*)。每个人称 *player* 或 *agent*. 每个人将为“避暑”这一目的选择或前往遮阳伞下或因遮阳伞下人太多而离开。由于你自己是一个人前来，你做出决策因此取决于人群整体的分布（在或不在遮阳伞下的概率分布），而受其他一个人的影响很小，毕竟你不认识他们，也就是你和他们是独立的。这种博弈就是本节要介绍的 *Mean field games*. 他是多人的博弈，更确切来讲是无限多 players 间的博弈，而 player 之间的相互作用很小可忽略，对于每个 player i 而言，其他 player 对其决策的影响是通过与其他 players 整体起作用的。假设共 N 个玩家，每个玩家的位置可以用一个 SDE 表示（仅建模，不一定符合实际）：

$$dX_t^i = \alpha^i(t, X_t, \bar{\mu}(t)) + dB_t^i, \forall i \in 1, 2, \cdots, N$$

其中，每个人的位置可以看作一个 Dirac 测度 $\delta_{X_t^i}$, 因此上述 $\bar{\mu}(t) = \frac{1}{N-1}\sum_{j\neq i}\delta_{X_t^j}$. $\alpha^i(t, X_t, \bar{\mu}(t))$ 为玩家 i 在时刻 t 和自己前当下位置 X_t^i 以及环境（其他的玩家的状态）$\bar{\mu}(t)$ 下做出的反馈、策略，或若把 X 当作位置，直接把 α 视作速度，$(B_t^i)_{t\geq 0}$ 为 Brownian Motion, 不同的玩家对应的布朗运动相互独立，当然每个玩家 i 也会有一个初始的位置 $X_0^i = x^i$ 这个也彼此独立。

”避暑”这一目的，相当于系统一个控制的目标，可定义一个成本函数：

$$J^i(\alpha^i) = \mathbb{E}\left[\int_0^T \left(f(X_t^i, \bar{\mu}_t) + \frac{1}{2}|\alpha_t|^2\right)dt + g(X_T^i, \bar{\mu}_t)\middle| X_0^i\right]$$

Remark:$J^i(\alpha^i)$ 并不是与其他玩家的策略无关，其他玩家的策略对玩家 i 的影响体现在 $\bar{\mu}_t$ 上。

既然是一个多人博弈，那就不得不提 *Nash Equilibrium.*

定义（Nash Equilibrium）：对于一个非合作博弈系统，假设有 N 个玩家，每个玩家的策略空间为 A_i, 称策略 $a = (a_1, a_2, \cdots, a_N)$ 为 Nash Equilibrium 策略，如果：

$$\forall i, \quad J^i(a_i, a_{-i}) \leq J^i(a_i', a_{-i}), \quad \forall a_i' \in A_i$$

其中符号 a_{-i} 表示 a 去掉第 i 个人后的其他人的策略.

Remark: 简而言之，Nash 均衡是指，单方面改变不能使得结果更优。从定义来看，Nash Equilibrium 并不是每个玩家的最优策略，只是相对其他玩家的一种 *compromise*, 其实要定义最优策略得要求:

$$\forall i, \quad J^i(a_i, a_{-i}') \leq J^i(a_i', a_{-i}'), \quad \forall a' \in A_1 \times A_2 \times \cdots \times A_N$$

纳什均衡可以由于是一种稳态，本质上可以看作不动点。做映射 $F : A_1 \times A_2, \cdots, A_N \to A_1 \times A_2, \cdots, A_N, a = (a_1, a_2, \cdots, a_N) \mapsto \prod_{i=1}^N \arg\min_{y_i \in A_i} J(y_i, a_{-i})$, 则 $F(a^*) = a^*$.

虽然 Nash 均衡对于每个玩家而言并不一定是最优的，但是对于整个系统而言是稳定的。也就是说当博弈系统处于纳什均衡是，系统达到稳态，不会产生变化，每个玩家的策略也将保持当前纳什均衡的策略。即：*No Incentive to deviate unilaterally from a compromise!*

因此，对于上述定义的博弈系统 (Mean Field Games), 每个玩家选择的最优策略将会是纳什均衡策略 $(\alpha_t^{1,*}\ \alpha_t^{2,*}, \cdots, \alpha_t^{N,*})$（你可以把它作为一种假设、前提或者我们期望的状态），这样系统的环境 $\bar{\mu}(t)$ 将会达到稳态。假设每个玩家的起始状态独立同分布，对应的布朗运动独立同分布，且有相同的成本函数，纳什均衡策略，那么每个玩家每个时刻的位置 $X_t^{i,*}$ 以及策略 $\alpha_t^{i,*}$ 将会是独立同分布 (一种对称性)，当玩家数量 $N \to +\infty$

时，根据大数定律，系统的环境抑或是针对于某个玩家 i, 看其他玩家整体的状态 $\bar{\mu}_t$ 将服从某一个理论分布，设这个分布为 m_t, 也就是

$$\bar{\mu}_t = \frac{1}{N-1}\sum_{j\neq i}\delta_{X_t^{j,*}} = \frac{1}{N}\sum_{j}\delta_{X_t^{j,*}} \to m_t, N \to +\infty$$

由于博弈系统处于纳什均衡状态，系统环境 m_t 处于稳态，玩家的策略也将会保持纳什均衡策略, 即相对于环境最优的策略。且各玩家满足独立同分布状态，我们以玩家 i 为代表来看。其在 Nash 均衡处满足的 SDE 变为

$$dX_t^{i,*} = \alpha^*(t, X_t^{i,*}, m_t)dt + dB_t^i$$

其中

$$\begin{cases} \alpha^* = \inf_\alpha J_m(\alpha) \\ m_t = \text{Law}(X_t^*) \end{cases}$$

可以按照如下解释进一步理解。假设系统环境为 m, 以此为输入通过方程组的方程 1，可以获得最优策略（Nash 均衡意义下），随后以这个最优策略为输入通过方程 2（与系统方程）更新下一次环境 m，纳什均衡意味着这两次的 m 具有一致性。这里由于玩家/粒子有微分方程，并不意味环境的稳态是随时间不变，而是上述的一致性。具体而言，一致性体现在方程组中的两个方程并不是独立的（先一后二、先二后一），而是两个方程相互耦合，即方程一中求最优策略依赖方程二的环境 m_t, 而二中更新环境依赖于一中的 α^*.

Remark: 令 $F : (\alpha, \mu) \mapsto (\arg\min_\alpha J_m(\alpha), \text{Law}(X_t^{\alpha^*}))$, 则 $F(\alpha^*, m) = (\alpha^*, m)$, 即 (α^*, m) 为一不动点。

求解上述系统可得（下述公式去掉星号，仍表示 Nash 均衡最优），可得最优策略为：$\alpha = -\partial_x u(x,t)$，$(m_t, u)$ 满足如下方程

$$\begin{cases} \left(\partial_t u + \frac{1}{2}\Delta u\right)(x,t) - \frac{1}{2}|\partial_x u(x,t)|^2 + f(x, m_t) = 0 \\ \partial_t m_t - \text{div}(m_t \partial_x u(x,t)) - \frac{1}{2}\Delta m_t = 0 \\ m_0 = m(0), \quad u(x,T) = g(x, m_T) \end{cases}$$

设 *Lagrange* 函数为 $L = T - V = \frac{1}{2}|\alpha_t|^2 + f(X_t, m_t)$, 其对应的 Hamilton 函数为 $H = T + V = \frac{1}{2}|\alpha_t|^2 - f(X_t, m_t) = \frac{1}{2}|\partial_x u(x,t)|^2 - f(X_t, m_t)$. 则 $\alpha = -D_p H(x, m, \nabla u)$, 相应的 MFGs 系统方程可写作：

$$\begin{cases} \left(\partial_t u + \frac{1}{2}\Delta u\right)(x,t)) - H(x, m, Du) = 0 \\ \partial_t m - \frac{1}{2}\Delta m - \text{div}\left(D_p H(x, m, Du)m\right) = 0 \\ m(0) = m_0, \quad u(x,T) = g(x, m(T)) \end{cases}$$

类似的可以得到更一般的 Mean Field Games(MFGs)。

设系统：$dX_t = B(X_t, \alpha_t, m_t)dt + \sqrt{2\nu}B_t$，$\nu$ 为常数或仅与时间 t 有关。成本函数：

$$J = \mathbb{E}\left[\int_0^T L(X_s, \alpha_s, m_s)ds + G(X_T, m(T))\middle| X_0\right]$$

其中 L 为 Lagrange 函数。设其对应的 Hamilton 函数为 $H = H(x, m, p)$。

则最优策略 α^* 将满足：$B(X_t, \alpha^*(x,t), m_t) = -D_pH(x, m, Du)$，最终系统 MFGs 系统方程变为：

$$\begin{cases} -\partial_t u - \nu\Delta u + H(x, m, Du) = 0 & \text{in } \mathbb{R}^d \times (0,T) \\ \partial_t m - \nu\Delta m - \operatorname{div}(D_pH(x, m, Du)\cdot m) = 0 & \text{in } \mathbb{R}^d \times (0,T) \\ m(0) = m_0, \quad u(x,T) = G(x, m(T)) \end{cases}$$

其中第一个方程为使成本函数最小化值函数 $u(t,x)$ 满足的带有终值时刻 T 条件的 “backward in time” 的 Hamilton-Jacobi-Bellman 方程，第二个方程为系统 SDE 带有初值条件的前向 Fokker-Planck 方程（又称 Kolmogorov Forward Equation）。H 为 L 的 Legendre 对偶 (或 Fenchel conjugate).

Remark: 在力学中，$L = \frac{p_i^2}{2m} - V(q_i,t)$,$H = \frac{p_i^2}{2m} + V(q_i,t) = p_i\dot{q}_i - L(q_i,p_i,t) = \frac{\partial L}{\partial \dot{q}_i}\dot{q}_i - L(q_i,p_i,t)$, 其中 q_i 为广义坐标，$p_i = \frac{\partial L}{\partial \dot{q}_i}$ 为广义动量，V 为势能,m 不同于 MFGs 的概率密度，这里为质量之意。因此，H 相对于广义动量 p_i(或 $\dot{q}_i$) 而言是 L 的 Legendre 对偶, 由于 L, H 相对 p_i 都为凸函数，L 亦为 H 的 Legendre 对偶。沿用上述记号，有：

$$\dot{p}_i = -\frac{\partial H}{\partial q_i} = -\frac{\partial V}{\partial q_i} \quad \dot{q}_i = \frac{\partial H}{\partial p_i} = \frac{p_i}{m}$$

需要指出，在 MFG 领域中与经典力学相反, 常取 $p = -\frac{\partial L}{\partial \dot{x}}$, 后续保持此设定, 因此 L 的 Legendre 对偶变为：

$$H(x, m, p) = \sup_{\dot{x}}[\frac{\partial L}{\partial \dot{x}}\dot{x} - L] = \sup_{\alpha}[-B(x, \alpha, m)\cdot p - L(x, m, \alpha)]$$

，此时力学中的 Hamilton 系统将变为：

$$\dot{\mathbf{x}} = -D_pH(\mathbf{p}, \mathbf{x}), \quad \dot{\mathbf{p}} = D_xH(\mathbf{p}, \mathbf{x})$$

这也就是为什么最优策略满足(忽略随机项)：$\dot{x} = B(X_t, \alpha^*(x,t), m_t) = -D_pH(x, m, Du)$。

我们将在下一节 MFGs 中的 Hamiton-Jacobi-Bellman 方程，本节剩下内容给出 MFGs 中 Fokker-Plank 方程。虽然 Fokker-Plank 方程在 Wasserstein 梯度流中已经出

现了，但是并没有给出关联 SDE 的随机过程 X_t 其概率密度 ρ_t 内在的 Fokker-Plank 方程。

给定高维 SDE:

$$dX_t = \mu(t, X_t)dt + \sigma(t, X_t)dB_t$$

其中 $\sigma(t, X_t)$ 假设为对角阵。设 $\rho(t)$ 为随机变量 X_t 的概率密度函数，取测试函数 $h : \mathbb{R}^d \to \mathbb{R}$，可微且满足 $h(-\infty) = h(+\infty) = h'(-\infty) = h'(\infty) = 0$.

根据 Ito 公式有：$dh(X_t) = (\nabla h\mu + \frac{1}{2}\Delta h\sigma^2)dt + \sigma\nabla h dB_t$

因此，$\mathbb{E}[dh(X_t)] = dt\mathbb{E}[\nabla h\mu + \frac{1}{2}\Delta h\sigma^2]$

左侧：$\mathbb{E}[dh(X_t)] = d\mathbb{E}[h(X_t)] = d < h, \rho_t >=< h, d\rho_t >$, 注意 $\frac{dh(x)}{dt} = 0$.

上述 “左侧” 等式有问题，一方面 $d(f(X_t))$ 本身只是记号没有任何含义，另一方面交换微分和期望（概率密度与微分 t 相关）也有问题。

右侧，利用分布积分（两次）可得：$dt\mathbb{E}[\nabla h\mu + \frac{1}{2}\Delta h\sigma^2] = dt < \frac{1}{2}\sigma^2\nabla\rho - \mathrm{div}(\mu\rho_t), h >$ 由于 h 的任意性，因此得到 Fokker-Plank 方程：

$$\partial_t\rho(t, x) = \frac{1}{2}\sigma^2\Delta\rho(t, x) - \mathrm{div}(\mu\rho(t, x))$$

上述形式没错，但是推导过程有问题，改正如下：

改正： $dh(X_t) = (\nabla h\mu + \frac{1}{2}\Delta h\sigma^2)dt + \sigma\nabla h dB_t$ 本质上对应：

$$h(X_t) = h(X_0) + \int_0^t (\nabla h\mu + \frac{1}{2}\Delta h\sigma^2)ds + \int_0^t \sigma\nabla h dB_s$$

左右两侧取期望：

$$\mathbb{E}[h(X_t)] = \mathbb{E}[h(X_0)] + \mathbb{E}\left[\int_0^t (\nabla h\mu + \frac{1}{2}\Delta h\sigma^2)ds\right] + 0$$

展开：

$$\int_\Omega h(x)\rho(t, x)dx = \mathbb{E}[h(X_0)] + \int_0^t \int_\Omega (\nabla h\mu + \frac{1}{2}\Delta h\sigma^2)\rho(s, x)dxds$$

等式两边对 t 取偏导，这时求偏导和积分可以正常交换次序

$$\int_\Omega h(x)\partial_t\rho(t, x)dx = \int_\Omega (\nabla h\mu + \frac{1}{2}\Delta h\sigma^2)\rho(t, x)dx$$

右侧做两次分布积分，得到：

$$< \nabla h\mu + \frac{1}{2}\Delta h\sigma^2, \rho >=< \frac{1}{2}\Delta(\sigma^2\rho) - \mathrm{div}(\mu\rho_t), h >$$

左侧为 $< \partial_t\rho(t, x), h >$, 由 h 的任一性有：

$$\partial_t\rho(t, x) = \frac{1}{2}\Delta(\sigma^2\rho) - \mathrm{div}(\mu\rho(t, x))$$

更一般的高维形式 (即 $\boldsymbol{\sigma}$ 不一定为对角阵)，假设

$$d\mathbf{X}_t = \boldsymbol{\mu}(\mathbf{X}_t, t)dt + \boldsymbol{\sigma}(\mathbf{X}_t, t)d\mathbf{W}_t$$

则:

$$\frac{\partial \rho(\mathbf{x}, t)}{\partial t} = -\sum_{i=1}^{N} \frac{\partial}{\partial x_i} \left[\mu_i(\mathbf{x}, t)\rho(\mathbf{x}, t)\right] + \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\partial^2}{\partial x_i \partial x_j} \left[D_{ij}(\mathbf{x}, t)\rho(\mathbf{x}, t)\right],$$

其中: $\mathbf{D} = \frac{1}{2}\boldsymbol{\sigma}\boldsymbol{\sigma}^{\mathsf{T}}$，分量形式为:

$$D_{ij}(\mathbf{x}, t) = \frac{1}{2}\sum_{k=1}^{M} \sigma_{ik}(\mathbf{x}, t)\sigma_{jk}(\mathbf{x}, t)$$

Example1 取 $dX = -\nabla V(X)dt + \sqrt{2}dt$, 则有: $\partial_t \rho(t, x) = \Delta\rho(t, x) + \mathrm{div}(\nabla V \rho(t, x))$

Example2 取 $dX_t = \alpha^*(t, X_t, m\rho_t)dt + dB_t$,$\alpha^*(t, X_t, \rho_t) = -\partial_x u(t, x)$ 则有:$\partial_t \rho(t, x) = \frac{1}{2}\Delta\rho(t, x) + \mathrm{div}(\partial_x u(t, x)\rho(t, x))$

Example3 取 $X_t = B(X_t, \alpha_t, m_t)dt + \sqrt{2\nu}B_t$,$B(X_t, \alpha^*(x, t), m_t) = -D_p H(x, \rho, Du)$ 则有: $\partial_t \rho(t, x) = \nu\Delta\rho(t, x) + \mathrm{div}(D_p H(x, \rho, Du)\rho(t, x))$

最后，给出系统方程满足上述 SDE（假设 σ 为对角阵):

$$dX_t = \mu(X_t, \alpha_t, m_t)dt + \sigma(X_t, \alpha_t, m_t)dB_t$$

的 MFGs:

$$\begin{cases} -\partial_t u(t, x) + H(t, x, \nabla u(t, x), D^2 u(t, x)) = 0 & \text{in } (0, T) \times \mathbb{R}^d \\ \partial_t m(t, x) = -\Delta \left[D_M H(t, x, Du, D^2 u) \cdot m\right] + \mathrm{div}\left[D_p H(x, \rho, Du, D^2 u) \cdot m\right] & \text{in } (0, T) \times \mathbb{R}^d \\ m(0) = m_0, \quad u(T, X_T) = G(X_T) & \text{in } \mathbb{R}^d. \end{cases}$$

其中: $H(t, x, p, M) := \sup_{\alpha \in A} -L(t, x, \alpha) - p \cdot b(t, x, \alpha) - \frac{1}{2}\mathrm{Tr}(\sigma\bar{\sigma}(t, x, \alpha)M)$, $D^2 u = (u_{x_i, x_j})$.

注:$H(t, x, p, M) = L(t, x, \alpha^*) - p \cdot b(t, x, \alpha^*) - \frac{1}{2}\mathrm{Tr}(\sigma\bar{\sigma}(t, x, \alpha^*)M)$, 对 M 求偏导有:

$$D_M H(t, x, p, M) = D_M \left(-\frac{1}{2}\mathrm{Tr}(\sigma\bar{\sigma}(t, x, \alpha^*)M)\right) = -\frac{1}{2}\sigma\bar{\sigma}(t, x, \alpha^*)$$

10.2 最优控制与 Hamilton-Jacobi-Bellman 方程

10.2.1 动态规划原理

现在，我们先考虑一个无随机版本的最优控制: $\dot{\mathbf{x}} = f(\mathbf{x}, \mathbf{u})$,$\mathbf{u}$ 为控制策略，$\mathbf{u} \in U$,U 为凸紧集。成本函数为: $J[\mathbf{u}; x, t] = \int_t^{t_T} L(\mathbf{x}, \mathbf{u})ds + G(\mathbf{x}(t_T))$, 初值 $\mathbf{x}(t) = x$

设值函数 $V(x,t)=\inf_{\mathbf{u}} J(\mathbf{u},x,t)$, 表示初始时刻为 t、粒子位置为 x 下的极小成本. 值函数满足下述定理。

定理（动态规划）： 设 $t_0 \le t \le t^{'} \le t_T$, 则：

$$V(x,t)=\inf_{\mathbf{u}}\left[\int_t^{t'} L(\mathbf{x}(s),\mathbf{u}(s),s)ds+V(y,t')\right]$$

其中 $y=x(t^{'}),\mathbf{x}(t)=x$.

这个定理就是平常做的离散版本动态规划的连续版本。先记右侧为 $\tilde{V}(t,x)$.

证明： 先证 $\tilde{V}(x,t)\le V(x,t)$. 由 $V(t,x)$ 的极小性，则固定 ϵ, 存在一个 a.e. 最优策略 $\mathbf{u}^{\epsilon}$, 其关联的轨迹记为 $\mathbf{x}^{\epsilon}$, 使得 $J(x,t;\mathbf{x}^{\epsilon})\le V(t,x)+\epsilon$. 另一方面，由 $\tilde{V}(t,x)$ 的极小性，有：

$$\tilde{V}(x,t)\le\int_t^{t'} L(\mathbf{x}^{\epsilon}(s),\mathbf{u}^{\epsilon}(s),s)ds+V(y,t')$$

且由 $V(y,t')$ 的极小性有：$V(y,t')\le J(y,t';\mathbf{u}^{\epsilon})$，因此

$$\tilde{V}(x,t)\le J(x,t;\mathbf{u}^{\epsilon})\le V(x,t)+\epsilon$$

对 $\epsilon\to 0$ 知，$\tilde{V}(x,t)\le V(x,t)$.

通过反证法证明另一方面，假设：$\tilde{V}(x,t)<V(x,t)$, 则对于任一 ϵ, 可取使得 $\tilde{V}(x,t)$ 趋近极小的策略 $\mathbf{u}^{\sharp}$, 设其对应的路径为 $\mathbf{x}^{\sharp}$, 使得

$$\int_t^{t'} L(\mathbf{x}^{\sharp}(s),\mathbf{u}^{\sharp}(s),s)ds+V(y,t')<V(x,t)-\epsilon$$

对后半段，根据 $V(y,t')$ 的极小性，可以选择策略 $\mathbf{u}^{\beta}$, 使得，$J(y,t';\mathbf{u}^{\beta})\le V(y,t')+\frac{\epsilon}{2}$. 因此可以构造出一个新的策略：

$$\begin{cases}\mathbf{u}^{\star}(s)=\mathbf{u}^{\sharp}(s)\ \text{for}\ s<t'\\ \mathbf{u}^{\star}(s)=\mathbf{u}^{\beta}\mathbf{u}(s)\ \text{for}\ t'<s.\end{cases}$$

则

$$\begin{aligned}V(x,t)-\epsilon>&\int_t^{t'} L(\mathbf{x}^{\sharp}(s),\mathbf{u}^{\sharp}(s),s)ds+V(y,t')\ge\\ &\ge\int_t^{t'} L(\mathbf{x}^{\sharp}(s),\mathbf{u}^{\sharp}(s),s)ds+J(y,t';\mathbf{u}^{\beta})-\frac{\epsilon}{2}=\\ &=J(x,t;\mathbf{u}^{\star})-\frac{\epsilon}{2}\ge V(x,t)-\frac{\epsilon}{2},\end{aligned}$$

推出矛盾。因此 $\tilde{V}(x,t)\ge V(x,t)$. 最终得 $V(x,t)=\inf_{\mathbf{u}}\left[\int_t^{t'} L(\mathbf{x}(s),\mathbf{u}(s),s)ds+V(y,t')\right]$ ■

Note: 上述推导并不要求最优控制策略 $\mathbf{u}^{*}$ 存在，也不要求值函数的可微性。

尽管上述动态规划原理展示了值函数的属性，但是并不好求解，我们希望推导出值函数满足的方程，这更能揭示值函数的本质。这个方程就是 Hamilton-Jacobi 方程：

$$-\partial_t V(x,t) + H(x, \nabla_x V(x,t)) = 0$$

这个方程就是无扩散/随机项的 MFGs 系统的 Hamilton-Jacobi-Bellman Equation。需要说明的是值函数 $V(t,x)$ 并不总是可微，若可微自然满足如上 PDE，若在某点处不可微（至少得要求连续），则在弱解意义下将满足如上 PDE，这个弱解意义就是“viscosity solution”！

下面，将基于上述动态规划原理结合一点点的变分法推导出值函数关联的 Hamilton-Jacobi Equation. 推导过程，假设值函数相对于 x,t 都可微，且最最优控制策略 $\mathbf{u}^*$ 存在。取 $t_0 \le t < t_T$，t 时刻位置 $\mathbf{x}(t) = x$，$t + h \le t_T$，根据动态规划原理有：

$$V(x,t) = \min_{\mathbf{u}\in U} \left[\int_t^{t+h} L(\mathbf{x}(s), \mathbf{u}(s)) ds + V\left(\mathbf{x}(t+h), t+h\right) \right]$$

根据链式求导，对第二项在 t 时刻处进行泰勒展开有：

$$\begin{aligned} &V(\mathbf{x}(t+h), t+h) = V(\mathbf{x}(t),t) + h\mathbf{x}'(t) \cdot \nabla_{\mathbf{x}} V(\mathbf{x}(t),t) + hV_t(\mathbf{x}(t),t) + o(h) \\ &= V(x,t) + hf(x,\mathbf{u}(t)) \cdot \nabla_{\mathbf{x}} V(x,t) + hV_t(x,t) + o(h). \end{aligned}$$

带入得：

$$V(x,t) = \min_{\mathbf{u}\in U} \left[\int_t^{t+h} L(\mathbf{x}(s), \mathbf{u}(s)) ds + V(x,t) + hf(x,\mathbf{u}(t)) \cdot \nabla_{\mathbf{x}} V(x,t) + hV_t(x,t) + o(h) \right]$$

化简得：

$$0 = hV_t(x,t) + o(h) + \min_{\mathbf{u}\in U} \left[\int_t^{t+h} L(\mathbf{x}(s), \mathbf{u}(s)) ds + hf(x,\mathbf{u}(t)) \cdot \nabla_{\mathbf{x}} V(x,t) \right]$$

令 $h \to 0^+$，等式两边同除以 h，由于此时：

$$\frac{1}{h} \int_t^{t+h} L(\mathbf{x}(s), \mathbf{u}(s)) ds = L(x, \mathbf{u}(t))$$

因此 $V_t(x,t) + \min\limits_{\mathbf{u}\in U} \left[L(x,\mathbf{u}(t)) + f(x,\mathbf{u}(t)) \cdot \nabla_{\mathbf{x}} V(x,t) \right] = 0, \forall x \in \mathbb{R}^d, t < t_T$

由于之前的定义：

$$H(x,p,m) = \sup_{\dot{x}} [\frac{\partial L}{\partial \dot{x}} \dot{x} - L] = \sup_{\mathbf{u}} [-f(x,\mathbf{u}) \cdot p - L(x,\mathbf{u})] = -\min_{\mathbf{u}} [f(x,\mathbf{u}) \cdot p + L(x,\mathbf{u})]$$

因此最终有 Hamilton Jacobi 方程成立：

$$-\partial_t V(x,t) + H(x, \nabla_x V(x,t)) = 0, \forall x \in \mathbb{R}^d, t < t_T$$

终止条件 $V(\mathbf{x},t_T)=G(\mathbf{x}(t_T))$. 且关联最优策略 $\mathbf{u}^*$ 的 $\mathbf{p}^*(t)=\nabla_{\mathbf{x}}V(\mathbf{x},t)$.

关于添加随机项的场景下，将得到 Hamilton-Jacobi-Bellman 方程，推导过程与上述类似，只不过是利用 Ito 公式链式求导。

对于随机过程 $X_s\omega:[t,T]\times\Omega\to\mathbb{R}^d$, 其初值 $X_t=x$，x 这里也是一随机变量。设其满足 SDE:

$$X_s^\alpha=x_0+\int_t^s b(r,X_r^\alpha,\alpha_t)dr+\int_t^s\sigma(r,X_r^\alpha,\alpha_r)dB_r$$

即：$dX_s^\alpha=b(s,X_s^\alpha,\alpha_t)ds+\sigma(s,X_s^\alpha,\alpha_s)dB_s$，这里假设 σ 为对角阵。

这里的上标 α 仅表示随机过程的轨道与 α 有关之意，其中 $(B_s)_{s\geq0}$ 为 N 维布朗运动 (起始时刻为 0 而非 t)，$\sigma:[0,T]\times\mathbb{R}^d\times A\to\mathbb{R}^{d\times N}$,$\alpha=(\alpha_s)$ 为控制过程，给定控制过程，则每个时刻 $\alpha_s\in A$,A 为控制变量的样本空间，而令 $\mathcal{A}$ 为控制过程（一个控制变量的序列）的集合，其在控制变量样本空间是正可测的。最优控制的目标为:

$$J(t,x,\alpha)=\mathbb{E}\left[\int_t^T L(s,X_s^\alpha,\alpha_s)ds+G(X_T)\middle|X_t=x\right]$$

引入值函数:$u(t,x)=\inf\limits_{\alpha\in\mathcal{A}}J(t,x,\alpha)$

同样的，值函数将有动态规划原理，对 $0\leq t\leq t'\leq T$, 有:

$$u(t,x)=\inf_{\alpha\in\mathcal{A}}\mathbb{E}\left[\int_t^{t'}L(s,X_s^\alpha,\alpha_s)ds+u(t',X_{t'}^\alpha)\middle|X_t=x\right]$$

假设值函数可微，类比上文，取 $t'=t+h<T,h>0$, 对可测函数（变量也是函数）$u(u(t+h,X_{t+h}^\alpha))$ 在时刻 t, 展开，注意这里要用 Ito 公式 x(下述只是一维的):

$$\begin{aligned}du(t,X_t)&=u_t(t,X_t)dt+u_x(t,X_t)dX_t+\frac{1}{2}u_{xx}(t,X_t)d^2X_t\\&=(u_t+u_xB+\frac{1}{2}\sigma^2)dt+u_x\sigma dW\end{aligned}$$

则

$$\begin{aligned}u(t,x)=\inf_{\alpha\in\mathcal{A}}\mathbb{E}\Bigg[&\int_t^{t+h}L(s,X_s^\alpha,\alpha_s)ds+u(t,x)+\int_t^{t+h}(\partial_tu(s,X_s^\alpha)+Du(s,X_s^\alpha)\cdot b(s,X_s^\alpha,\alpha_s)\\&+\frac{1}{2}\mathrm{Tr}(\sigma\bar{\sigma}(s,X_s^\alpha,\alpha_s)D^2u(s,X_s^\alpha)))ds\Bigg].\end{aligned}$$

注意 dW 服从高斯分布,$\mathbb{E}[\int fdW]=0$。$\bar{\sigma}$ 为 σ 的共轭转置，$D^2u=(u_{x_i,x_j})$.

令 $h\to0^+$, 并等式两边同除 h, 得:

$$0=\partial_tu(t,x)+\inf_{\alpha\in A}\left[L(t,x,\alpha)+Du(t,x)\cdot b(t,x,\alpha)+\frac{1}{2}\mathrm{Tr}(\sigma\sigma^*(t,x,\alpha)D^2u(t,x))\right]$$

按照之前的记号，Hamiton 函数:

$$H(t,x,p,M):=\sup_{\alpha\in A}\left[-L(t,x,\alpha)-p\cdot b(t,x,\alpha)-\frac{1}{2}\mathrm{Tr}(\sigma\bar{\sigma}(t,x,\alpha)M)\right]$$

则得到带终值条件的 Hamilton-Jacobi-Bellman 方程：

$$\begin{cases} -\partial_t u(t,x) + H(t,x,\nabla u(t,x), D^2 u(t,x)) = 0 \quad \text{in } (0,T)\times\mathbb{R}^d \\ u(T,X_T) = G(X_T) \quad \text{in } \mathbb{R}^d. \end{cases}$$

特别的，若 $\sigma(s, X_s^\alpha, \alpha_s) = \sqrt{2}\nu$,$\nu$ 为一常数或仅与时间 t 有关，则可取 Hamilton 函数为 $H(t,x,p) = \sup_{\alpha\in A}[-L(t,x,\alpha) - p\cdot b(t,x,\alpha)]$

得到之前得到的 MFGs 中的 Hamilton-Jacobi-Bellman 方程：

$$\begin{cases} -\partial_t u(t,x) - \nu\Delta u(t,x) + H(t,x,\nabla u(t,x)) = 0 \quad \text{in } (0,T)\times\mathbb{R}^d \\ u(T,X_T) = G(X_T) \quad \text{in } \mathbb{R}^d. \end{cases}$$

假设最优控制过程 α^* 的存在性，则：$H(t,x,p) = L(t,x,\alpha^*) - p\cdot b(t,x,\alpha^*)$, 因此有：$p^* = \nabla u$ 且 $b(t,x,\alpha^*) = -H_p(t,x,\nabla u)$ ■.

上述，我们从假设值函数可微、最优控制 $\alpha^* \in \mathcal{A}$ 存在出发，通过值函数定义经动态规划原理推出了值函数满足的 H-J 方程。那反过来，能否通过 H-J 方程得到值函数呢?

我们先一点一点看，仍假设值函数可微、最优控制 $\alpha^*\mathcal{A}$ 存在。记 $\alpha^*(t,x) \in A$ 满足 Hamiton 函数定义的那个 α, 且 H 中 $p = \nabla u$, $M = \Delta u$. 我们可以通过每个时刻的 $\alpha^*(t,x)$ 和 HJ 方程**构造出最优控制过程** α^*，这称之为 **Verification Theorem**。

定理（Verification Theorem）：假设最优控制 $\alpha^* \in \mathcal{A}$ 存在。设 $(X_s^{\alpha^*})$ 为下述随机微分方程的解：$dX_s^{\alpha^*} = b(s, X_s^{\alpha^*}, \alpha_s^*)ds + \sigma(s, X_s^{\alpha^*}, \alpha_s^*)dB_s$

其中令 $\alpha_s^* = \alpha^*(s, X_s^{\alpha^*}) \in A$. $\alpha^*(s, X_s^{\alpha^*})$ 为每个时刻 Hamilton 函数定义所关联的极大值点，即通过每个时刻的最优反馈，构造了每个时刻最优控制 α_s^*，最终形成最优控制过程 α^*。则值函数 $u(t,x)$ 满足

$$u(t,x) = J(t,x,\alpha^*)$$

简单的说明如下：根据 Ito 公式有：

$$\begin{aligned} G(X_T^{\alpha^*}) = u(T, X_T^{\alpha^*}) = u(t,x) + \int_t^T (\partial_t u(s, X_s^{\alpha^*}) + Du(s, X_s^{\alpha^*})\cdot b(s, X_s^{\alpha^*}, \alpha_s^*) \\ + \frac{1}{2}\mathrm{Tr}(\sigma\sigma^*(s, X_s^{\alpha^*}, \alpha_s^*)D^2u(s, X_s^{\alpha^*})))ds + \int_t^T \sigma^*(s, X_s^{\alpha^*}, \alpha_s^*)Du(s, X_s^{\alpha^*})\cdot dB_s \end{aligned}$$

根据 H 的定义有：

$$\begin{aligned} H(t,x,Du(t,x),D^2u(t,x)) = -L(t,x,\alpha^*(t,x)) - Du(t,x)\cdot b(t,x,\alpha^*(t,x)) \\ - \frac{1}{2}\mathrm{Tr}(\sigma\sigma^*(t,x,\alpha^*(t,x))D^2u(t,x)) \end{aligned}$$

带入上式并取期望有：

$$\begin{aligned}\mathbb{E}\left[G(X_T^{\alpha^*})\right] &= u(t,x) + \mathbb{E}\left[\int_t^T (\partial_t u(s,X_s^{\alpha^*}) - H(s,X_s^{\alpha^*}, Du(s,X_s^{\alpha^*}), D^2u(s,X_s^{\alpha^*})) - L(s,X_s^{\alpha^*},\alpha_s^*))ds\right] \\ &= u(t,x) - \mathbb{E}\left[\int_t^T L(s,X_s^{\alpha^*},\alpha_s^*)ds\right].\end{aligned}$$

其中，最后一个等式是带入了 u, H 的 Hamilton-Jacobi 方程. 即得：

$$u(t,x) = \mathbb{E}\left[\int_t^T L(s,X_s^{\alpha^*},\alpha_s^*)ds + G(X_T^{\alpha^*})\right] = J(t,x,\alpha^*)$$

这就验证了上述构造的 α^* 为最优控制过程。 ■

那再进一步问，如果一个函数是 HJ 方程的经典解，那这个解是不是就是值函数呢? 答案显然是不一定的。如下我们给出两个例子，一个 HJ 方程的解就是值函数，另一个例子说明几乎处处满足 HJ 方程的解不唯一，因此不是所有 Lipschitz 连续的解都是值函数。

Example1（HJ 方程的解为值函数）： 设 $L(x,u) \in C^1$ 为 Lagrange 函数，且相对于速度变量/控制变量 u 严格凸，设 $\dot{x} = f(x,u)$ 中 f 满足 Lipschitz 条件：$|f(x,u) - f(y,u)| \le C|x-y|$, 且 f 为线性结构：$f(x,u) = A(x)u + B(x)$, 相应的 A, B 都是 Lipschitz 的.H 为 L 如上述定义的广义 Legendre 变换。设 $\Phi(x,t)$ 为 Hamiton-Jacobi 方程的解且 $\Phi(x,T) = G(x_T)$, 则 $\Phi(x,t) = V(x,t), \forall t \in [0,T]$. 其中 $V(x,t)$ 为上述定义的值函数.

证明见Calculus of Variations and Partial Differential Equations Page193-194.

Example2 (HJ 方程的解不为值函数): 假设 HJ 方程为：$-\partial_t V(t,x) + |\partial_x V(t,x)|^2 = 0$, 且终止成本为 0(设终止时间为 1)，即 $V(1,x) = 0$. 显然 $V_1 \equiv 0$ 是上述带终值约束的 HJ 方程的解（本质是这个问题的值函数，为粘性解），同样：

$$V(x,t) = \begin{cases} 0 \quad \text{if} \quad |x| \ge 1-t \\ |x| - 1 + t \quad \text{if} \quad |x| < 1-t \end{cases}$$

也是上述带终值约束的 HJ 方程的几乎处处可微的解（本质非粘性解，因此非值函数）。也就是 HJ 方程满足 Lipschitz 连续的解不唯一，而值函数只有一个，因此 HJ 方程的解不一定是值函数。

上述全部的讨论都在值函数 u 可微，最优控制过程 α^* 存在两个假设条件上。显然真实情况没有那么乐观。（值函数是通过下确界定义出来的，不存在存不存在一说，即使最优控制过程不存在，值函数也可正常定义。）

关于**确定性版本的最优控制的存在性问题**请查看Calculus of Variations and Partial Differential Equations的 Sec4.8 （Page 202-214），他给出了最优控制问题存在性的一个充分条件，当然在这个条件下，值函数亦是 Lipschitz 连续的。

关于**值函数 u 的可微性**，参考Zhiping Rao 最优控制与动态规划 ，可以给出值函数 Lipschitz 连续（几乎处处可微）的一个充分条件。

10.2.2 Viscosity Solution

以确定性版本的最优控制为例，介绍引入 H-J 方程 *viscosity solution* 的动机。

上一节我们知道假设值函数 $V(t,x)$ 在点 (t,x) 处可微，则他将满足 H-J 方程：

$$-\partial V(t,x)+H(t,x\partial_x V)=0$$

给出的例子和资料告诉我们：若终值函数 G 为Lipschitz 连续（> 绝对连续（牛顿-莱布尼兹成立的充要条件）> 一致连续 > 连续），且 Lagrange 函数 L 满足一定条件时（如相对于速度的凸性、Lipschitz 连续性，见上一节末的参考资料），或其他充分条件下，值函数 $V(t,x)$ 是 Lipschitz 连续的。由 Rademacher 定理——Lipschitz 函数几乎处处可微——可得值函数 $V(t,x)$ 将几乎处处满足 H-J 方程。然而，上一节最后的列子也告诉我们，几乎处处满足 H-J 方程的 Lipschitz 连续函数并不一定是值函数。

那问题是我们将问题划归为求解 Hamilton-Jacobi 方程，他可能有很多几乎处处满足方程的解（Lipschitz 连续的解），那么那个解是值函数呢？我们需要在几乎处处满足 HJ 方程外，另加一些条件，使得解具有额外的更好性质，如唯一性。

在本节，我们首先通过一种正则化的 HJ 方程的解，并令正则项系数趋于 0（这个过程叫做 *vanishing viscosity*）的方式推导出上述正则化解的极限（正则化系数为 0）满足的性质（有关测试函数及测试函数满足的不等关系）。并由这个性质定义出 HJ 方程的一类弱解——粘性解（*Viscosity Solution*）。当然，我们会给出经典解和粘性解的一致性（一方面任意可微的 HJ 方程经典解就是粘性解。另一方面 HJ 的粘性解若在某点 (t,x) 处可微，则在该点满足 HJ 方程。作为其中的一个特例有，任一足够光滑的粘性解就是经典解）以表明定义的合理性。随后我们给出上述定义的粘性解的唯一性（至多一个粘性解）原理，在对 Hamilton 函数的某些假设下。最后给出值函数就是最优控制问题关联的 HJ 方程的粘性解（也是唯一粘性解）这一结论。请注意，粘性解只是 HJ 方程的弱解，在某些约束下（如各种 Lipschitz 连续约束）下，粘性解（值函数）几乎处处满足 HJ 方程，但是并不是处处满足，因此不一定是 HJ 方程的经典解。当然若值函数（在定义的区域上）处处可微，正如前文我们所推导的那样，它将处处满足 HJ 方程，则其就是 HJ 的经典解。另外，由于在某些条件成立下，粘性解唯一，而值函数为粘性解，上述正则化 HJ 的解的极限也天然满足粘性解的定义，因此我们可以根据正则化的 HJ 方程（的解），从处处满足 HJ 方程的众多候选 Lipschitz 函数中选择出值函数。

10.3 More on Hamilton Jacobi Equation

前一节推导的倒向 HJ 方程:

$$\begin{cases} -\partial_t V(x,t) + H(x, \nabla_x V(x,t)) = 0, \forall x \in \mathbb{R}^d, \quad \text{in } (0,T) \times \mathbb{R}^d \\ u(T, X_T) = G(X_T) \quad \text{in } \mathbb{R}^d. \end{cases}$$

取 $u(t,x) = V(T-t,x)$, 则变为带初值条件的 HJ 方程:

$$\begin{cases} \partial_t u(x,t) + H(x, \nabla_x u(x,t)) = 0, \forall x \in \mathbb{R}^d, \quad \text{in } (0,T) \times \mathbb{R}^d \\ u(0, X_0) = G(X_0) \quad \text{in } \mathbb{R}^d. \end{cases}$$

本小节将补充关于带初值约束的 HJ 方程更多内容，包括特征方程，以及一个简单情况下 ($H = H(p)$ 仅显式与 p 变量有关，对应 $L = L(v)$ 仅于速度变量有关，且两为凸函数) 的具体案例。更多内容请参考Evans 的 PDE第 3、10 章以及Calculus of Variations and Partial Differential Equations第四章。

10.3.1 Characteristic Method

介绍其思想，特征方程，关于一些初值问题为啥可能会不存在光滑解（用例子），举 $L \equiv L(v)$ 且凸的例子，展示 Characteristic Method 的三个方程每一项是什么，然后就是此种情况下有个 Hopf_Lax 公式。

10.3.2 Simple Case——Hopf-Lax 公式

10.3.3 Viscosity Solution

10.4 Variational Mean Field Games 与 OT

11 Stochastic Optimal Transport

参考Stochastic Optimal Transportation;Yongxin Chen的博士论文From Schrödinger bridges to Optimal Mass Transport及其他综述性质的文章; 以及Stochastic optimal transport and Hamilton-Jacobi-Bellman equations

11.1 Stochastic Optimal Control 与 OT

11.2 Schrödinger's Problem and Schrödinger Bridge

11.3 Sinkhorn-Style Solvers for Diffusion Models

参考：OT tutorial 2023 bunne, 不过，我们将在 Sec.14再展开。

12 Monge Ampere Equation

13 Sampling Method

本章主要基于Sinho Chewi的书Log-Concave Sampling和ETH-401-4634-24L: Diffusion Models, Sampling and Stochastic Localization前几次课程。主要集中于 Langevin Algorithm. 以关联下一章以及前面的内容。

14 Advanced Generated Model——Diffusion and Flow

最优传输与当下时兴的生成模型关联密切，毕竟最优传输的动态视角（Benamou-Brenier 定理）给出了连接初始分布（如高斯噪音）到目标分布的最优的概率路径！这正可以为 Diffusion or Flow 给一个前向过程，且这个路径是能量的最小！

本章主要聚焦于最近的生成模型（选材都是“经典”且高引的）：

DDPM 及一般的 Score based Model, 并选择介绍DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps这篇文章。

Flow Matching方法与更一般的Generator Matching: Generative modeling with arbitrary Markov processes。也会介绍条件生成的Improving and generalizing flow-based generative models with minibatch optimal transport和最近的Analyzing and Improving Optimal Transport for Conditional Flow-Based Generation, 以及源域目标域耦合非独立相关的Stochastic interpolants with data-dependent couplings, 其前身为一个可视为统一 flow 与 diffusion 的框架Stochastic Interpolants。

NeurIPS 2024 Tutorial——Flow Matching for Generative Modeling(讲稿参见 Flow Matching Guide and Code) 是很好的 flow matching 视频教程, 里面还涉及一些一般流形上的 flow matching （Flow Matching on General Geometries）以及对称 (Equivariant

flow matching) 不变性, 这里就不展开了。当然 Score based model 也可以在流形上做如Riemannian Score-Based Generative Modelling, 这里也不展开。

Rectified Flow:Flow Straight and Fast与Rectified Flow: A Marginal Preserving Approach to Optimal Transport, 以及最近的Rectified Diffusion: Straightness Is Not Your Need in Rectified Flow

最后介绍一些桥相关的方法如: Likelihood Training of Schrödinger Bridge using Forward-Backward SDEs Theory, Diffusion Schrödinger Bridge with Applications to Score-Based Generative Modeling, Image-to-Image Schrodinger Bridge以及Let us Build Bridges: Understanding and Extending Diffusion Generative Models。推荐一个上述文章作者Qiang Liu的一个网站Let us Flow Together

补充一下最近的一个何凯明团队的Mean Flows for One-step Generative Modeling

A 附录 A：Calculus of Variations

变分法 1：Euler-Lagrange，变分法 2：Hamilton.

A.1 Euler-Lagrange Equation

A.2 一些简单案例

A.2.1 等周不等式的多种解法

Parseval 等式

A.2.2 悬链线

A.2.3 高斯分布——均值、方差固定下的最大熵

A.3 Geodesic

A.3.1 平面上的测地线

A.3.2 球面上的测地线

A.3.3 一般黎曼度量下的测地线

A.4 拉普拉斯方程与 Poisson 方程

A.5 最优控制的变分求解

不含不确定性的手写笔记见变分法的两个手写笔记末页或最优控制变分法与庞特里亚金极小值原理，Calculus of Variations and Partial Differential Equations 第 4 章

A.6 变分法与 OT、梯度流

B 附录 B：Stochastic Differential Equation

UCB Lawrence C. Evans 的小册子，这是个人注记版：An Introduction to Stochastic Differential Equations (Lawrence C. Evans) 含注记。

B.1 Brown Motion

B.2 Ito 微积分

B.3 一般线性 SDE 的求解

B.4 Diffusion Process

B.5 Langevin Equation and Sampling

视采样为优化问题 - 以 Langevin Diffusion 为例

B.6 Feynman-Kac Formula

SDE 与 PDE 的一种关联：一个偏微分方程的解可以写成关于一个随机过程泛函的期望值，称这个表达式为偏微分方程的概率表示，即 *Feynman-Kac* 公式。推导的参考资料：一个简要介绍-无推导，带 Cauchy 边值下的费曼-卡茨公式推导粗略、基于鞅的推导。

B.7 Kolmogorov 前向后向方程

一个很详细的前后向方程推导：基石背后的数学：详细推导 Kolmogorov 方程 前向方程（FP）方程的一个另外的推导什么是 Fokker-Planck 方程? 后向方程的另一推导利用 Feynman-Kac Formula

B.8 动态规划下的最优控制：Hamilton-Jacobi-Bellman 方程

有随机与无随机版详细推导、一个 SDE 下的不严谨推导过程、一个 ODE 下（无随机）的不严谨推导

参考文献

[1] Filippo Santambrogio. Optimal transport for applied mathematicians. Springer International Publishing, 2015.

[2] Bruno Levy and Erica Schwindt. Notions of optimal transport theory and how to implement them on a computer. 2017.

[3] Jiajin Li, Jianheng Tang, Lemin Kong, Huikang Liu, Jia Li, Anthony Man-Cho So, and Jose Blanchet. A convergent single-loop algorithm for relaxation of gromov-wasserstein in graph data. 2023.

[4] Xinyu Ma, Xu Chu, Yasha Wang, Yang Lin, Junfeng Zhao, Liantao Ma, and Wenwu Zhu. Fused gromov-wasserstein graph mixup for graph-level classifications. 2023.

[5] Laetitia Chapel, Mokhtar Z Alaya, and Gilles Gasso. Partial optimal tranport with applications on positive-unlabeled learning. Advances in Neural Information Processing Systems, 33:2903–2913, 2020.

[6] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. Mathematics of Computation, 87(314):2563–2609, 2018.

[7] Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo Rohde. Generalized sliced wasserstein distances. Advances in neural information processing systems, 32, 2019.

[8] Samir Chowdhury, David Miller, and Tom Needham. Quantized gromov-wasserstein. 2021.

[9] Alessio Figalli and Federico Glaudo. An invitation to optimal transport, Wasserstein distances, and gradient flows. 2021.

[10] L. Ambrosio. Gradient Flows in Metric Spaces and in the Space of Probability Measures. Gradient flows : in metric spaces and in the space of probability measures, 2008.