生成对抗网络的几何观点以及算法优化

项目研究成果

最优传输理论与其几何基础

项目编号：201910055055.

负 责 人：崔嘉珩.

所在院系：数学科学学院 统计学.

联系电话：18622705306.

指导教师：吴春林.

摘　要

为了解决在生成对抗网络（GAN）中出现的梯度消失问题，人们提出了许多方法，引人关注的 WGAN 就是其中之一。它采用了最优传输理论中的 Wasserstein 距离作为判别器的损失函数。通过最优传输中的 Kantorovich 方法，我们可以知道在最优传输的观点下，生成对抗网络的生成器 G 中的映射等价于最优传输问题中的最优传输映射；生成对抗网络的判别器 D 中的 Wasserstein 距离等价于最优传输问题中的 Kantorovich 势函数。

我们还可以从凸几何角度来考虑 WGAN。若用 Brenier 方法，我们可以证明并且在代价函数为凸的情况下，最优传输映射和 Wasserstein 距离可以互相推导。同时，在代价函数为常见的欧氏距离的情况下，Kantorovich 方法和 Brenier 方法是等价的，这让我们可以在几何的观点下理解最优传输问题。

而在得到这些结论之前，我们需要一些基本的数学概念。微分流形是微分几何中的一个重要分支，我们将用微分流形的知识来介绍与描述上述内容。

关键词**:** 微分流形；最优传输；Kantorovich 方法；Brenier 方法

Abstract

To solve the gradient vanishing problem that arises in adversarial generative networks (GAN), many methods have been proposed, and the attention-grabbing WGAN is one of them. It adopts the Wasserstein distance in optimal transportation theory as the loss function of the discriminator. By Kantorovich's approach in optimal transport, we can know that the mapping in the generator G is equivalent to the optimal transport map in the optimal transport problem; the Wasserstein distance in the discriminator D is equivalent to the Kantorovich potential function.

We can also consider WGAN from the point of view of convex geometry. If we use Brenier's approach, we can prove and derive the optimal transportation map and Wasserstein distance from each other in the case where the cost function is convex. Also, the Kantorovich's approach and the Brenier's approach are equivalent in the case where the cost function is the Euclidean distance, which allows us to understand the optimal transportation problem in a geometric point of view.

Before we get to these conclusions, we need some basic mathematical concepts. Differentiable manifold is an important branch of differential geometry. We will use the knowledge of differentiable manifolds to introduce and describe the WGAN theory.

Key Words: Differentiable Manifold; Optimal Transportation Theory; Kantorovich's Approach; Brenier's Approach

目 录

目录

第一章　微分流形理论

微分流形是微分几何中的一个重要分支，是微分几何的主要研究对象。现在的深度学习模型一般都涉及到高维数据，这些数据直观上很可能难以解释。对此，机器学习理论中有如下的“流形假设”：高维数据的特征实际上位于低维空间内的一个流形之上。

为了介绍相关的理论，我们在本章会介绍流形的概念，并简单介绍深度学习与流形的关系。

第一节　流形的概念

Definition 1. 流形 *(Manifold)*[1]：*n* 维流形 Σ 是一个拓扑空间，由一组开集 $\Sigma \subset \cup_\alpha U_\alpha$ 覆盖。对于每个开集 U_α，都有一个同胚 $\varphi_\alpha : U_\alpha \to \mathbb{R}^n$，配对 $(U_\alpha, \varphi_\alpha)$ 组成一个图 *(chart)*。图的并集形成一个图集 *(atlas)*$\mathcal{A} = \{(U_\alpha, \varphi_\alpha)\}$。如果 $U_\alpha \cap U_\beta \neq \emptyset$，则图传输映射由 $\varphi_{\alpha\beta} : \varphi_\alpha (U_\alpha \cap U_\beta) \to \varphi_\beta (U_\alpha \cap U_\beta)$ 给出，

$$\varphi_{\alpha\beta} := \varphi_\beta \circ \varphi_\alpha^{-1} \tag{1.1}$$

设 $\mathcal{X}$ 是环境空间 (ambient space)，μ 是定义在 $\mathcal{X}$ 上的概率分布，表示为密度函数 $\mu : \mathcal{X} \to \mathbb{R}_{\geq 0}$。$\mu$ 的支撑集 $\Sigma(\mu) := \{\mathrm{x} \in \mathcal{X} | \mu(x) > 0\}$ 是一个低维流形。

设 $(U_\alpha, \varphi_\beta)$ 是一个局部图 (local chart)，$\varphi_\alpha : U_\alpha \to \mathcal{Z}$ 称为编码映射，参数域 $\mathcal{Z}$ 称为隐空间 (latent space) 或特征空间 (feature space)，其维度远低于 $\mathcal{X}$ 的维度。点 $\mathrm{x} \in \Sigma$被称为样本，其参数 $\varphi_\alpha(\mathrm{x})$ 被称为x的代码或特征。φ_α 的逆映射 $\psi_\alpha := \varphi_\alpha^{-1} : \mathcal{Z} \to \Sigma$称为解码映射，其给出了流形的局部参数表示。

此外，编码映射 $\varphi_\alpha : U_\alpha \to \mathcal{Z}$ 能够诱导出一在特征空间 $\mathcal{Z}$ 上定义的前推 (push-forward) 概率测度 $(\varphi_\alpha)^* \mu$。“前推”指的是：对于任何可测集 $B \subset \mathcal{Z}$，都有

$$(\varphi_\alpha)^* \mu(B) := \mu\left(\varphi_\alpha^{-1}(B)\right) \tag{1.2}$$

深度学习模型（如 GAN、自编码器等）的目标是学习编码映射 φ_α、解码映射 $\psi_\alpha : \mathcal{Z} \to \Sigma$（即流形的参数表示），以及前推概率 $(\varphi_\alpha)_* \mu$ 等。下面我们将解释自编码器 (AutoEncoder) 如何学习流形和其上的分布。

第二节　深度学习中的流形假设

深度学习是解决许多机器学习任务的主流技术，包括计算机视觉、自然语言处理等。它在各个领域的性能均优于传统机器学习方法，并取得了巨大的成功。同时，大部

分业界人士都认为奠定深度学习的理论基础，尤其是解释性理论至关重要。不幸的是，深度学习的解释性理论发展仍不完善，且解释深度学习模型一般是非常困难的。我们希望用流形来解释深度学习模型，我们认为解释深度学习需要依靠数据中的流形结构。

在深度学习界存在一个公认的流形假设[2]：自然的数据在环境空间 $\mathcal{X}$ 中，其是极高维的空间。同样类型的数据（如人脸图片）在 $\mathcal{X}$ 中的一个流形 Σ 之上，其服从一个概率分布 μ。流形假设还认为：同样类型的数据的特征集中在特征空间 $\mathcal{Z}$（其维度比 $\mathcal{X}$ 低很多）中的一个非线性低维流形附近，特征向量的分布为 ν。各种深度学习方法的主要焦点是从实际数据中学习流形结构并获得流形的参数表示。

例如，自编码器 [6] 学习编码映射 $\varphi_\theta : \mathcal{X} \to \mathcal{Z}$ 和解码映射 $\psi_\theta : \mathcal{Z} \to \mathcal{X}$。输入流形 Σ 的参数表示由解码映射 ψ_θ 给出。重构的流形 $\tilde{\Sigma} = \psi_\theta \circ \varphi_\theta(\Sigma)$ 近似于输入流形。我们将在本章第三节着重介绍自编码器的流形解释。

此外，深度学习模型还学习和控制由在特征空间上定义的编码器 $\tilde{\Sigma} = \psi_\theta \circ \varphi_\theta(\Sigma)$ 引起的分布。一旦得到参数流形结构，它就可以作为生成模型在 $\tilde{\Sigma}$ 上随机生成一个样本 [5]；它还可以进行图像去噪 [7] [8]，图像去噪的几何解释为：编码再解码会将噪声样本投影到表示干净图像流形的 $\tilde{\Sigma}$ 上的最近点，这样就能给出一张和该样本最相近的去噪图像。这可以作为流形假设成立的一个证明。

第三节　流形、自编码器与生成模型

自编码器通常用于无监督学习 [6]，它被应用于压缩、去噪、预训练等。从几何角度来说，自编码器学习数据的低维特征，将其表示为参数多面体流形，即从特征空间（参数域）到环境空间的分段线性 (Piecewise Linear，简称为 PL) 映射，PL 映射的像是一 $\mathcal{X}$ 中的高维多面体流形。然后自动编码器利用该多面体流形作为原分布流形在各种应用中的数据逼近。

1.3.1　自编码器的流形解释

从结构上讲，自编码器是一种前馈的、非递归的神经网络，其输出层与输入层具有相同的节点数，其目的是重建自身的输入。自编码器的隐藏层为一瓶颈层 (bottleneck)，其维度显著低于输入层和输出层，其功能是降维。自编码器的输入空间是环境空间 $\mathcal{X}$，输出空间也是环境空间 $\mathcal{X}$。瓶颈层的输出空间是特征空间 $\mathcal{Z}$。

自编码器通常由编码器 (encoder)φ 和解码器 (decoder)ψ 两部分组成。编码器获取一个样本 $\mathrm{x} \in \mathcal{X}$ 并将其映射到 $z \in \mathcal{Z}, z = \varphi(x)$，图像 z 通常被称为 x 的潜在表示 (latent representation)。编码器 $\varphi : \mathcal{X} \to \mathcal{Z}$ 将 Σ 映射到它的潜在表示 $D = \varphi(\Sigma)$。之后，解码器 $\psi : \mathcal{Z} \to \mathcal{X}$ 将 z 映射到与 x 维度相同的重构 $\tilde{\mathbf{x}}$，$\tilde{x} = \psi(z) = \psi \circ \varphi(x)$。

为了希望 x 与 $\tilde{x}$ 尽量接近，自编码器需要通过训练以最小化重建误差：

$$\varphi, \psi = \text{argmin}_{\varphi,\psi} \int_{\mathcal{X}} \mathcal{L}(x, \tilde{x}) d\mu(\mathrm{x}) \tag{1.3}$$

其中 $\mathcal{L}(\cdot,\cdot)$ 是损失函数，例如平方误差。重构的流形 $\tilde{\Sigma} = \psi \circ \varphi(\Sigma)$ 则可被视为 Σ 的近似。

1.3.2 自编码器的实际应用

在实际应用中，情况通常如下所示：编码器和解码器都是以 ReLU DNN 的形式实现的，由 θ 参数化。记 $X = \left\{\mathrm{x}^{(1)}, \mathrm{x}^{(2)}, \ldots, \mathrm{x}^{(k)}\right\}$ 为训练数据集，$X \subset \Sigma$。自编码器优化 L^2 损失函数：

$$\min_{\theta} \mathcal{L}(\theta) = \min_{\theta} \frac{1}{k} \sum_{i=1}^{k} \left\| \mathrm{x}^{(i)} - \psi_\theta \circ \varphi_\theta \left(\mathrm{x}^{(i)}\right) \right\|^2 \tag{1.4}$$

其中编码器 φ_θ 和解码器 ψ_θ 都是 PL 映射。编码器 φ_θ 包含环境空间的单元分解 $\mathcal{D}(\varphi_\theta)$[?]：

$$\mathcal{D}(\varphi_\theta) : \mathcal{X} = \bigcup_{\alpha} U_\theta^\alpha \tag{1.5}$$

其中 U_θ^α 是凸多面体，其上对 φ_θ 的限制 (restriction) 是仿射映射。类似地，分段线性映射 $\psi_\theta \circ \varphi_\theta$ 包含多面体单元分解 $\mathcal{D}\left(\psi_\theta, \varphi_\theta\right)$，这是 $\mathcal{D}(\varphi_\theta)$ 的细化 (refinement)。重建的多面体流形具有参数表示 $\psi_\theta : \mathcal{Z} \to \mathcal{X}$，它近似于数据中的流形 Σ。

1.3.3 用自编码器进行图片生成

假设 $\mathcal{X}$ 是所有 $n \times n$ 个彩色图像的空间，其中每个点代表一个图像。我们可以定义一个概率测度 μ，它表示图像表示要学习内容的概率。这里用人脸举例，人脸的形状是由有限数量的基因决定的。人脸照片是由人脸的几何结构、光线、摄像机参数等决定的。因此，假设所有的人脸照片都集中在一个有限维流形上是合理的，我们称之为人脸照片流形 Σ。

通过使用大量真实的人脸照片，我们可以训练一个自动编码器来学习人脸照片流形。学习过程产生解码映射 $\psi : \mathcal{Z} \to \tilde{\Sigma}$，即重构流形的参数表示。我们从一个简单分布（如正态分布）中随机生成一个向量 $z \in \mathcal{Z}$，再用正态分布到人脸特征分布的最优传输算法 T 把它投射到 $\mathcal{Z}$ 上，得到 x，则 x 应距离人脸特征分布流形较近，$\varphi(x) \in \tilde{\Sigma}$ 会以大概率给出一个人脸图像，而该图像很可能与训练集中的图像均不相同。因此这可以作为生成人脸照片的生成模型。

具体到我们的实验中，我们所采用的例子是手写阿拉伯数字 [9]。比起人脸数据集，阿拉伯数字的特征更少。相同数字之间的共性大于手写的随机性，同时随机因素也较

少，这样在训练上更简单，同时我们可以通过比较相同训练波数 (epoch) 下传统 GAN 和我们算法的生成结果，来体现出模型的优劣。在手写数字上完成优化后，我们尝试对人脸数据集进行训练，得到了一些不错的生成图片。我们的算法以及结果详细可见项目结题报告。

第二章　最优传输理论

为了解决 GAN[10] 中出现的梯度消失问题，人们提出了 WGAN[11]。它采用了最优传输理论中的 Wasserstein 距离作为判别器的损失函数，这样当生成的样本分布和训练集的样本分布的支撑集没有相交部分时，利用 Wasserstein 距离作为判别器的损失函数的 WGAN 就总可以得到一个用来优化生成器的合适梯度。

我们在这章介绍一些最优传输理论的基础概念，包括 Wasserstein 距离、Kantorovich 势与 Kantorovich 对偶问题，并给出求离散最优传输映射的算法——Sinkhorn 算法[12]。

第一节　基本概念与 Kantorovich 问题

Definition 2. $Polish$ 空间[13][14][15][16]：完备，可分的度量空间 (X,d)。

Definition 3. 给定 $Polish$ 空间 (X,d)，用 $\mathcal{P}(X)$ 表示 $Polish$ 空间 (X,d) 上面所有 $Borel$ 概率测度的集合。

Definition 4. 前推映射 *(push-forward mapping)*T 与前推算子 *(push-forward operator)*T^**(*或记为 $T_{\#}$*)*[13]：假设 X,Y 为两个 $Polish$ 空间，$T:X\to Y$ 是一个 $Borel$ 映射，$\mu\in\mathcal{P}(X)$ 是一个测度，则可以定义 Y 上的测度 $\nu=T^*\mu$：

$$T^*\mu(E)=\mu(T^{-1}(E)),\quad \forall E\subset Y,\ E\ Borel \tag{2.1}$$

我们称 T 将 μ 前推到了 ν。

2.1.1　Monge 最优传输问题与最优传输映射

最优传输理论起源于法数学家蒙日 (Monge) 在 1781 年提出的 Monge 最优传输问题[18]：

Definition 5. *Monge* 问题：$\mu\in\mathcal{P}(X),\nu\in\mathcal{P}(X)$：

$$\begin{aligned}&\min_T \int_X c(x,T(x))d\mu(x)\\ &s.t.\quad T^*\mu=\nu\end{aligned} \tag{2.2}$$

最优传输映射：*Monge* 问题 *(2.2)* 的解 T 称为从 μ 到 ν 的最优传输映射。

我们在这里通俗地解释一下最优传输问题[19]：

厨房需要给一个办公室进行小吃的分配。厨房里面有几种小吃 (小吃的经验分布为 μ)，并且知道每种小吃的总份数；也知道办公室里每个同事的饭量 (饭量的经验分布为

ν)（这里小吃种类和同事人数不一定相等）。每一位同事对每一种小吃都有一个偏好，如给其喜欢的小吃 1 个单位，其对厨房的总评分便 +1；一般，则评分为 0；不喜欢，则评分为-1。这个偏好可以总结为一个函数，即偏好函数，由于 Monge 问题是求最小值，我们可以令代价函数 $c(x, y)$ 为偏好函数的相反数。

Monge 问题即为：我们要将所有的小吃都分配出去（假设可以分配的份数为任意实数），这个分配方法 T 需要使得整个办公室对厨房的总评分最高。因此 T 为最优传输映射。

T 除了使代价最小外，还是保测度的。这是因为我们的分配要保证把小吃都分配出去，且从厨房拿出去的小吃总量等于办公室收到的小吃总量。

值得庆幸的是，常见的最优传输问题都与此问题类似，且这种问题是一定可解的，我们会在第三节中介绍如何求解该类问题。

2.1.2 最优传输的 Kantorovich 问题

Definition 6. 传输计划 *(transportation plan)*：一个 $X \times Y$ 上的联合测度 $\gamma(A \times B)$，其描述从 A 被分到 B 的份数。

Definition 7. $\mu \in \mathcal{P}(X)$，$\nu \in \mathcal{P}(Y)$，给定代价函数 $c(x, y)$。$\mathcal{A}_{opt}(\mu, \nu)$ 为所有满足 $\gamma(A \times Y) = \mu(A), \gamma(X \times B) = \nu(B)$ 的最优传输计划的全体 *(*即满足 $\pi_x^*\gamma = \mu_\theta, \pi_y^*\gamma = \nu$ 的概率测度 $\gamma \in \mathcal{P}(X \times Y)$ 的全体*)*，则最优传输的 *Kantorovich* 问题 *(*又称 *Kantorovich* 公式*)* 为：

$$\min_{\gamma \in \mathcal{A}_{opt}} \int_{X \times Y} c(x, y) d\gamma(x, y) \tag{2.3}$$

我们给出如下定理，具体证明请见 [4]：

Theorem 1. 当 $c(x, y)$ 是连续的，并且满足 $\mu(x) \neq 0, \forall x \in X$ 时有：

$$\inf(\textit{Monge} \text{ 问题}) = \min(\textit{Kantorovich} \text{ 问题}) \tag{2.4}$$

由此可以知道 *Kantorovich* 问题可以看成 *Monge* 问题的一个松弛 *(relaxation)*。

Kantorovich 问题的一些优势[14][15][16]：

- $\mathcal{A}$ 是非空的。
- $\mathcal{A}$ 是紧的和凸的。
- 最优传输计划 γ 包含了最优传输映射 T。
- 对于大多数 $c(x, y)$，Kantorovich 问题的解是存在的，而 Monge 问题不一定。

2.1.3 $\gamma \in \mathcal{A}_{opt}$ 的充要条件

以下的 c 均代表代价函数 $c(x, y)$:

Definition 8. *c*-循环单调 *(c-cyclical monotone)*[14]：称 $\Gamma \subset X \times Y$ 是 *c*-循环单调的，如果满足条件：对于 $(x_i, y_i) \in \Gamma, 1 \leq i \leq N$，下式对于所有置换 σ 成立：

$$\sum_{i=1}^{N} c\left(x_i, y_i\right) \leq \sum_{i=1}^{N} c\left(x_i, y_{\sigma(i)}\right) \tag{2.5}$$

Definition 9. *c*-变换 *(c-transform)*[4]*:*

- $\psi : Y \to \mathcal{R} \bigcup \{\pm\infty\}$ 为任意一个函数，则它的 c_+-变换 $\psi^{c+} : X \to \mathcal{R} \bigcup \{-\infty\}$ 定义为

$$\psi^{c+}(x) := \inf_{y \in Y} c(x, y) - \psi(y) \tag{2.6}$$

- $\psi : Y \to \mathcal{R} \bigcup \{\pm\infty\}$ 为任意一个函数，则它的 c_--变换 $\psi^{c-} : X \to \mathcal{R} \bigcup \{+\infty\}$ 定义为

$$\psi^{c-}(x) := \sup_{y \in Y} -c(x, y) - \psi(y) \tag{2.7}$$

- $\varphi : X \to \mathcal{R} \bigcup \{\pm\infty\}$ 为任意一个函数，则它的 c_+-变换 $\psi^{c+} : Y \to \mathcal{R} \bigcup \{-\infty\}$ 定义为

$$\varphi^{c+}(x) := \inf_{y \in Y} c(x, y) - \varphi(y) \tag{2.8}$$

- $\varphi : X \to \mathcal{R} \bigcup \{\pm\infty\}$ 为任意一个函数，则它的 c_--变换 $\psi^{c-} : Y \to \mathcal{R} \bigcup \{+\infty\}$ 定义为

$$\varphi^{c-}(x) := \sup_{y \in Y} -c(x, y) - \varphi(y) \tag{2.9}$$

Definition 10. *c*-凹 *(c-concave)* 和 *c*-凸 *(c-convex)*[14] [15] [16]：

- 我们称 $\varphi : X \to \mathcal{R} \bigcup \{-\infty\}$ 是 *c-concave* 的，如果存在 $\psi : Y \to \mathcal{R} \bigcup \{-\infty\}$，使得 $\varphi = \psi^{C+}$
- 我们称 $\varphi : X \to \mathcal{R} \bigcup \{+\infty\}$ 是 *c-convex* 的，如果存在 $\psi : Y \to \mathcal{R} \bigcup \{+\infty\}$，使得 $\varphi = \psi^{C-}$
- 我们称 $\psi : Y \to \mathcal{R} \bigcup \{-\infty\}$ 是 *c-concave* 的，如果存在 $\varphi : X \to \mathcal{R} \bigcup \{-\infty\}$，使得 $\psi = \varphi^{C+}$
- 我们称 $\psi : Y \to \mathcal{R} \bigcup \{+\infty\}$ 是 *c-convex* 的，如果存在 $\varphi : X \to \mathcal{R} \bigcup \{+\infty\}$，使得 $\psi = \varphi^{C-}$

Definition 11. *c*-上微分 *(c-superdifferential)* 和 *c*-下微分 *(c-subdifferential)*[14] [15] [16] [17]：

- 令 $\varphi : X \to \mathcal{R}\bigcup\{-\infty\}$ 是 *c-concave* 函数，则它的 *c*-上微分 $\partial^{c+}\varphi \subset X \times Y$ 定义为

$$\partial^{c+}\varphi := \{(x,y) \in X \times Y : \varphi(x) + \varphi^{c+}(y) = c(x,y)\} \tag{2.10}$$

其中

$$\partial^{c+}\varphi(x) := \left\{y|(x,y) \in \partial^{c+}\varphi\right\} \tag{2.11}$$

- 令 $\varphi : X \to \mathcal{R}\bigcup\{+\infty\}$ 是 *c-convex* 函数，则它的 *c*-下微分 $\partial^{c-}\varphi \subset X \times Y$ 定义为

$$\partial^{c-}\varphi := \{(x,y) \in X \times Y : \varphi(x) + \varphi^{c-}(y) = -c(x,y)\} \tag{2.12}$$

其中

$$\partial^{c-}\varphi(x) := \left\{y|(x,y) \in \partial^{c-}\varphi\right\} \tag{2.13}$$

Theorem 2. 最优传输基本定理[17]：$c : X \times Y \to \mathcal{R}$ 是连续的，$\mu \in \mathcal{P}(X), \nu \in \mathcal{P}(Y)$，若 $c(x,y)$ 满足 $c(x,y) \le a(x) + b(y)$，其中 $a(x) \in L^1(\mu), b(y) \in L^1(\nu)$，则以下条件等价：

- $\gamma \in \mathcal{A}_{opt}$
- $supp(\gamma)$ 是 c-循环单调的
- 存在 c-concave 函数 φ 使得 $\max\{\varphi, 0\} \in L^1$ 并且 $supp(\gamma) \subset \partial^{c+}\varphi$

根据最优传输基本定理，$\forall\gamma \in \mathcal{A}_{opt}$，一定存在 c-concave 函数 φ 使得 $supp(\gamma) \subset \partial^{c+}\varphi$，这保证了我们之后解 Kantorovich 问题的方法是理论上可行的。

第二节　Kantorovich 对偶

2.2.1　Kantorovich 对偶问题

注意到，Kantorovich 问题实际上是一个线性规划问题，而一切线性规划问题都存在它的一个对偶问题。

Definition 12. 令 $\mu \in \mathcal{P}(x), \nu \in \mathcal{P}(Y)$，则 *Kantorovich* 问题的对偶问题为[4]：

$$\begin{aligned} &\max_{\varphi\in L^1(\mu),\psi\in L^1(\nu)} \{\int_X \varphi(x)d\mu(x) + \int_Y \psi(y)d\nu(y)\} \\ &s.t. \quad \varphi(x) + \psi(y) \le c(x,y) \end{aligned} \tag{2.14}$$

2.2.2 Kantorovich 对偶的解释

实际上我们可以说明以下等式成立：

$$\inf_{\gamma\in\mathcal{A}(\mu,\nu)}\int_{X\times Y}c(x,y)d\gamma(x,y)=\sup_{\varphi,\psi}\{\int_X\varphi(x)d\mu(x)+\int_Y\psi(y)d\nu(y)\} \tag{2.15}$$

首先我们注意到

$$\sup_{\varphi,\psi}\{\int_X\varphi d\mu+\int_Y\psi d\nu\}-\int_{X\times Y}(\varphi(x)+\psi(y))d\gamma=\begin{cases}0 & \text{if } \gamma\in\mathcal{A}(\mu,\nu)\\ +\infty & \text{otherwise}\end{cases} \tag{2.16}$$

从而我们可以去掉限制 $\gamma\in\mathcal{A}(\mu,\nu)$，化简为：

$$\min_{\gamma}\int_{X\times Y}cd\gamma+\sup_{\varphi,\psi}\{\int_X\varphi d\mu+\int_Y\psi d\nu\}-\int_{X\times Y}(\varphi(x)+\psi(y))d\gamma \tag{2.17}$$

合并一些项即为：

$$\sup_{\varphi,\psi}\{\int_X\varphi d\mu+\int_Y\psi d\nu\}+\inf_{\gamma}\int_{X\times Y}(c(x,y)-(\varphi(x)+\psi(y)))d\gamma \tag{2.18}$$

由于

$$\inf_{\gamma}\int_{X\times Y}(c(x,y)-(\varphi(x)+\psi(y)))d\gamma=\begin{cases}0 & \text{if } \varphi(x)+\psi(y)\le c(x,y) \text{ on } X\times Y\\ -\infty & \text{otherwise}\end{cases} \tag{2.19}$$

所以 $\min_{\gamma\in\mathcal{A}}\int_{X\times Y}c(x,y)d\gamma(x,y)$ 的对偶问题为 $\max_{\varphi+\psi\le c}\{\int_X\varphi(x)d\mu(x)+\int_Y\psi(y)d\nu(y)\}$
由此我们知道该式成立

$$\inf_{\gamma\in\mathcal{A}(\mu,\nu)}\int_{X\times Y}c(x,y)d\gamma(x,y)=\sup_{\varphi,\psi}\{\int_X\varphi(x)d\mu(x)+\int_Y\psi(y)d\nu(y)\} \tag{2.20}$$

利用 c-concave 函数函数定义以及最优传输基本定理，我们有：

$$\begin{aligned}\inf_{\gamma}\int c(x,y)d\gamma(x,y)&=\sup_{\varphi}\{\int\varphi(x)+\varphi^{c+}(y)d\gamma(x,y)\}\\&=\sup_{\varphi}\{\int\varphi(x)d\mu(x)+\int\varphi^{c+}(y)d\nu(y)\}\end{aligned} \tag{2.21}$$

Theorem 3. 对偶定理：如果对偶问题的最大值可以被取到，则 (φ,ψ) 可以表示成 (φ,φ^{C+}) 形式，其中 φ 为 *c-concave* 函数。

Definition 13. *Kantorovich* 势函数：如果 (φ,φ^{C+}) 使得对偶问题达到最值，则称 *c-concave*

函数 φ 为该问题的 *Kantorovich* 势函数。

第三节　求解离散最优传输问题的方法

2.3.1　离散最优传输问题

在实际应用中，我们考虑的是离散形式的 Kantorovich 问题，也就是说我们要考虑的是从一个离散概率分布向量到另一个离散概率分布向量的最优传输问题。下面先介绍向量间的耦合矩阵。

Definition 14. 耦合矩阵集合 *(coupling matrix set)*[17]：

$$U(r,c) = \{P \in \mathcal{R}_+^{n*m} | P1_m = r, P^T 1_n = c\} \tag{2.22}$$

这里的 r,c 指分布向量，$U(r,c)$ 包含了所有从 c 传输到 r 的传输方案。

Definition 15. 给定分布向量 r，c，代价矩阵 M，则离散的 *Kantorovich* 问题为：

$$d_M(r,c) = \min_{P \in U(r,c)} P \odot M == \min_{P \in U(r,c)} \sum P_{ij} M_{ij} \tag{2.23}$$

$d_M(r,c)$ 被称作从 r 到 c 的 *Wasserstein* 距离 *(*或 *Earth-Mover* 距离，简称为 *EM* 距离*)*。P 被称作最优传输矩阵。

我们要做的就是求解最优传输矩阵 P，其与最优传输映射 T 在 r 与 c 维数相等时是等价的，详见我们的结题报告。由上述定义可见，最优传输映射和 Wasserstein 距离是等价的，这与第三章第二节的结论是完全一致的。

2.3.2　Sinkhorn 算法

回到分配小吃的问题：假设我们有五种小吃，每种小吃的份数分布为 $[4,2,6,4,4]$，则其分布向量 $c = [0.2, 0.1, 0.3, 0.2, 0.2]$；

有八个人，其胃口的分布向量为 $r = [0.15, 0.15, 0.15, 0.2, 0.1, 0.1, 0.1, 0.05]$；

每个同事对每种小吃的喜好程度矩阵为 M，M 给定。

则所有满足每列之和为 c，且每行之和为 r 的矩阵都属于 $U(r,c)$，其中有且仅有一个矩阵 P 能达到 Wasserstein 距离，我们可以用 Sinkhorn 算法来求出其近似值。

Sinkhorn 算法有无正则化和有正则化两种不同形式，其中正则化[21] 会让分配更加趋向于平均分配[17]。我们在结项报告中的生成算法只需要采用无正则化的 Sinkhorn 算法，使用的是 Python 中 POT 包[20] 的 ot.emd() 函数，其算法为运筹学中非常基础的单纯形法，在此并不赘述，具体算法请见 [22]。

由 2.1.2 节和 2.1.3 节，我们知道离散 Kantorovich 问题一定有解。Sinkhorn 算法证明[12]：无论是否有正则化问题，都可以通过迭代的方法求解 P 的近似值。这是因为在满足原问题的最优传输条件以及保测度条件后，可以把原问题转化成一组属于矩阵放缩的数学问题的等式，每一步分别满足一个等式，最终迭代一定会收敛。

第三章　凸几何角度理解

第一节　最优传输映射的存在性

我们知道 Kantorovich 问题可以看成 Monge 问题的一个松弛[14]。根据定理 1，当 $c(x,y)$ 是连续的，且满足 $\mu(x)=0,\forall x\in X$ 时，有 inf(Monge 问题) = min(Kantorovich 问题)。

但对于最优传输问题来讲，他们并不是总有解的。我们可以证明，若 $c(x,y)=h(x-y)$，h 是严格凸函数，则一定存在最优传输映射。而对于一般的代价函数，如 L^2 代价，其是严格凸函数，即最优传输映射一定存在。我们给出以下两个定理[4]：

Theorem 4. 假设 $c(x,y)\in C^1(X,Y)$，φ 为相应的 *Kantorovich* 势函数，$(x_0,y_0)\in supp(\gamma)$，则 $\nabla\varphi(x_0)=\nabla_x c(x_0,y_o)$

Theorem 5. 给定 μ 和 ν 为紧区域 $\Omega\subset\mathbb{R}^d$ 上的概率测度，则当 h 是严格凸函数的时候一定存在最优传输映射 T，并且满足 $T(x)=x-(\nabla h)^{-1}(\nabla\varphi(x))$

第二节　最优传输映射与 Kantorovich 势函数的等价性

定理 5 说明了 h 是严格凸函数时，最优传输映射 T 和 Kantorovich 势函数 $\varphi(x)$ 是等价的。在这里我们给出定理 5 的证明：

证明：假设 ρ 是满足条件 $\pi_{x\#}\rho=\mu,\pi_{y\#}\rho=\nu$ 的联合分布，任取一个在 ρ 的支撑上的点 (x_0,y_0)，由定义 $\varphi^c(y_0)=\inf\limits_x\{c(x,y_0)-\varphi(x)\}$，因此

$$\nabla\varphi(x_0)=\nabla_x c(x_0,y_0)=\nabla h(x_0-y_0), \tag{3.1}$$

因为 h 是严格凸的，因此 ∇ 是可逆的.

$$x_0-y_0=(\nabla h)^{-1}(\nabla\varphi(x_0)), \tag{3.2}$$

因此 $y_0=x_0-(\nabla h)^{-1}(\nabla\varphi(x_0))$.

第三节　L^2 代价情况下 Kantorovich 方法与 Brenier 方法的等价性

Kantorovich 方法是从最优传输理论对进行 WGAN 的数学理论进行解释；同时，我们也可以从凸几何方向对其进行解释，这引出了 Brenier 方法。在这里我们先对 Kantorovich 方法以及 Brenier 方法进行说明。然后我们给出在完全离散的情况进行证明两种方法在 L^2 代价函数下具有等价性，并推广到半连续情况。这让我们可以在几何的观点下理解最优传输问题。

假设 μ 有在 X 上的紧支撑 Ω，Ω 是 X 上的凸域.

$$\Omega = supp(\mu) = \{x \in X | \mu(x) > 0\} \tag{3.3}$$

用狄拉克度量 $\nu = \sum_{j=1}^{k} \nu_j \epsilon(y - y_j)$ 将空间 Y 离散为 $Y = \{y_1, y_2, ..., y_k\}$，其总量满足

$$\int_{\Omega} d\mu(x) = \sum_{i=1}^{k} v_i \tag{3.4}$$

3.3.1 Kantorovich 方法

我们定义离散的 Kantorovich 势[4] $\varphi : Y \to \mathbb{R}, \varphi(y_j) = \varphi_j$, 则有

$$\int_{Y} \varphi d\nu = \sum_{i=1}^{k} \varphi_j v_j \tag{3.5}$$

ϕ 的 c-变换由下式给出：

$$\varphi^c(x) = \min_{1 \le j \le k} \{c(x, y_j) - \varphi_j\} \tag{3.6}$$

这包括一个 X 的胞腔剖分 (cell decomposition)：

$$X = \bigcup_{i=1}^{k} W_i(\varphi) \tag{3.7}$$

每个胞腔有

$$W_i(\varphi) = \{x \in X | c(x, y_i) - \phi_i \le c(x, y_i) - \phi_j, \forall 1 \le j \le k\}. \tag{3.8}$$

根据 Wasserstein 距离的定义（即 Kantorovich 对偶问题 (2.14) 的最大值[4]）和公式 (3.5)，我们可以如下定义能量：

$$E(\varphi) = \int_{X} \varphi^c d\mu + \int_{Y} \varphi d\nu \tag{3.9}$$

然后我们可以得到

$$E_D(\varphi) = \sum_{i=1}^{k} \varphi_i \left(\nu_i - w_i(\varphi)\right) + \sum_{j=1}^{k} \int_{W_j(\varphi)} c\left(x, y_j\right) d\mu \tag{3.10}$$

这里的 $w_i(\varphi)$ 是胞腔 $W_i(\varphi)$ 的度量，即

$$w_i(\varphi) = \mu\left(W_i(\varphi)\right) = \int_{W_i(\varphi)} d\mu(x) \tag{3.11}$$

则 μ 和 ν 间的 Wasserstein 距离有

$$W_c(\mu, \nu) = \max_{\varphi} E(\varphi) \tag{3.12}$$

3.3.2 Brenier 方法

Kantorovich 的对偶方法适用于普通的代价函数。当成本函数是 L^2 范数下的距离 $c(x, y) = |x - y|^2$ 时，我们可以直接应用 Brenier 方法[23].

我们定义一个高度向量 (height vector)$h = (h_1, h_2, ..., h_k) \in \mathbb{R}^n$，其由 k 个实数组成。对于每个 $y_i \in Y$，我们构造一个在 X，$\pi_i(h) :< x, y_i > +h_i = 0$ 上定义的超平面。我们将 Brenier 势函数定义为

$$u_h(x) = \max_{i=1}^{k} \{\langle x, y_i \rangle + h_i\} \tag{3.13}$$

则 $u_h(x)$ 是一个凸函数. $u_h(x)$ 的图是具有支撑平面 $\pi_i(h)$ 的无限凸多面体.

图的投影诱导 Ω 的一个多边形划分 (polygonal partition)

$$\Omega = \bigcup_{i=1}^{k} W_i(h) \tag{3.14}$$

其中每个胞腔 $W_i(h)$ 是 u_h 在 Ω 上一个面的投影。

$$W_i(h) = \{x \in X \mid \nabla u_h(x) - y_i\} \cap \Omega \tag{3.15}$$

$W_i(h)$ 的度量由下式给出：

$$w_i(h) = \int_{W_i(h)} d\mu \tag{3.16}$$

这样，在每个胞腔 $W_i(h)$ 上的凸函数 u_h 都是线性函数 $\pi_i(h)$ ，因此，梯度映射

$$\nabla u_h : W_i(h) \to y_i, i - 1, 2, \cdots, k \tag{3.17}$$

将每个 $W_i(h)$ 映射到单个点 y_i。根据 Alexandrov 定理[24] 和 Gu-Luo-Yau 定理[3]，我们得出以下结论：

Theorem 6. 设 Ω 是 $\mathbb{R}^n$ 中的紧凸域，$\{y_1, ..., y_k\}$ 是 $\mathbb{R}^n$ 中的一组不同点，μ 是 Ω 上的概率度量。对于任何 $\nu = \sum_{i=1}^{k} \nu_i \delta_{y_i}$，且 $\sum_{i=1}^{k} \nu_i = \mu(\Omega)$，存在 $h = (h_1, h_2, ..., h_k) \in \mathbb{R}^k$，在

不考虑常数 $(c, c, ..., c)$ 的情况下，有 $w_i(h) = \nu_i$ 对于所有的 i 都成立。向量 h 是凹函数

$$E_B(h) = \sum_{i=1}^{k} h_i \nu_i - \int_0^h \sum_{i=1}^{k} w_i(\eta) d\eta_i \tag{3.18}$$

的最大点.

进一步，∇u_h 在所有 μ 到 ν 的传输映射 $T_{\#}\mu = \nu$ 中，能够最小化 L^2 代价函数

$$\int_\Omega |x - T(x)|^2 d\mu \tag{3.19}$$

3.3.3 Kantorovich 方法与 Brenier 方法的等价性

在离散情况下，当 $c(x, y) = \frac{1}{2}|x - y|^2$ 时，根据定理 5，我们有：

$$T(x) = x - \nabla\phi(x) = \nabla(\frac{x^2}{2} - \varphi(x)) = \nabla u(x) \tag{3.20}$$

在这种情况下，Brenier 势能 $u_h(x)$ 和 Kantorovich 势能 $\varphi(x)$ 有如下关系：

$$u_h(x) = \frac{1}{2}|x|^2 - \varphi(x) \tag{3.21}$$

在半连续情况下，(3.21) 式仍然成立，具体证明较为复杂，涉及到 Minkowski 问题和 Alexandrov 问题，在此不展开赘述，具体证明可见 [3]，[4]。

参考文献

[1] B. Riemann (1867).

[2] Yin H. (2008) Learning Nonlinear Principal Manifolds by Self-Organising Maps. In: Gorban A.N., Kégl B., Wunsch D.C., Zinovyev A.Y. (eds) Principal Manifolds for Data Visualization and Dimension Reduction. Lecture Notes in Computational Science and Enginee, vol 58. Springer, Berlin, Heidelberg.

[3] Xianfeng Gu, Feng Luo, Jian Sun, and Shing-Tung Yau. Variational principles for minkowski type problems, discrete optimal transport, and discrete monge-ampere equations. Asian Journal of Mathematics (AJM), 20(2):383 C 398, 2016.

[4] Lei, N., Su, K., Cui, L., Yau, S.-T., and Xianfeng Gu, D., A Geometric View of Optimal Transportation and Generative Model, arXiv e-prints, 2017.

[5] Lei, N., Luo, Z., Yau, S.-T., and Xianfeng Gu, D., Geometric Understanding of Deep Learning, arXiv e-prints, 2018.

[6] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell., 35(8):1798–1828, August 2013.

[7] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composingn robust features with denoising autoencoders. In Proceedings of the 25th International Conference on Machine Learning, ICML ’08, pages 1096–1103, New York, NY, USA, 2008. ACM.

[8] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. J. Mach. Learn. Res., 11:3371–3408, December 2010.

[9] Yann LeCun, THE MNIST DATABASE of handwritten digits. Courant Institute, NYU Corinna Cortes, Google Labs, New York Christopher J.C. Burges, Microsoft Research, Redmond.

[10] Goodfellow, I. J., Generative Adversarial Networks, arXiv e-prints, 2014.

[11] Martin Arjovsky, Soumith Chintala, and L´eon Bottou. Wasserstein GAN. In International Conference on Machine Learning, pages 214–223, 2017.

[12] Sinkhorn, R.. (1964). A relationship between arbitrary positive matrices and doubly stochastic matrices. Ann. Math. Statist. 35, 876–879. doi:10.1214/aoms/1177703591

[13] Peyré, G. and Cuturi, M., Computational Optimal Transport, arXiv e-prints, 2018.

[14] Santambrogio, F.. (2015) Optimal Transport for Applied Mathematicians. Progress in Nonlinear Differential Equations and Their Applications vol 87. Springer, Berlin, Heidelberg.

[15] Villani, C.. (2008) Optimal Transport-Old and New. Springer Berlin, Heidelberg, NewYork, HongKong, London, Milan, Paris, Tokyo.

[16] Yann Ollivier, Hervé Pajot, Cedric Villani. Optimal Transportation Theory and Applications. Cambridge University Press.

[17] Bruno Lévy, Erica L. Schwindt. Notions of optimal transport theory and how to implement them on a computer. Computers and Graphics, Volume 72, 2018, Pages 135-148, ISSN 0097-8493, https://doi.org/10.1016/j.cag.2018.01.009.

[18] G. Monge. Mémoire sur la théorie des déblais et des remblais. Histoire de l’ Académie Royale des Sciences de Paris, avec les Mémoires de Mathématique et de Physique pour la même année, pages 666–704, 1781.

[19] Stock, M.. Website: https://michielstock.github.io/posts/2017/2017-11-5-OptimalTransport/, 2017

[20] Rémi Flamary and Nicolas Courty, POT Python Optimal Transport library, Website: https://pythonot.github.io/, 2017

[21] Cuturi, M. (2013) Sinkhorn distances: lightspeed computation of optimal transportation distances.

[22] HILLIER, F. S., AND LIEBERMAN, G. J. 1990. Introduction to Operations Research, 5th ed. McGraw-Hill.

[23] Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. Comm. Pure Appl. Math., 44(4):375–417, 1991.

[24] A. D. Alexandrov. Convex polyhedra Translated from the 1950 Russian edition by N. S. Dairbekov, S. S. Kutateladze and A. B. Sossinsky. Springer Monographs in Mathematics. Springer-Verlag, Berlin, 2005.