Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Artificial Intelligence—Feature Article

深度学习的几何学解释

雷娜 [a,#], 安东生 [b,#], 郭洋 [b], 苏科华 [c], 刘世霞 [d], 罗钟铉 [a], 丘成桐 [e], 顾险峰 [b,e,*]

[a] DUT-RU Co-Research Center of Advanced ICT for Active Life, Dalian University of Technology, Dalian 116620, China
[b] Department of Computer Science, Stony Brook University, Stony Brook, NY 11794-2424, USA
[c] School of Computer Science, Wuhan University, Wuhan 430072, China
[d] School of Software, Tsinghua University, Beijing 100084, China
[e] Center of Mathematical Sciences and Applications, Harvard University, Cambridge, MA 02138, USA

ARTICLE INFO

Article history:
Received 2 March 2019
Revised 31 August 2019
Accepted 11 September 2019
Available online 11 January 2020

关键词
生成
对抗
深度学习
最优传输
模式崩溃

摘要

本文从几何角度来理解深度学习，特别是提出了生成对抗网络（GAN）的最优传输（OT）观点。自然数据集具有内在的模式，该模式可被概括为流形分布原理，即同一类高维数据分布于低维流形附近。GAN主要完成流形学习和概率分布变换两项任务。其中，后者可以用经典的OT方法来实现。从OT的角度来看，生成器用于计算OT映射，而判别器用于计算生成数据分布与真实数据分布之间的Wasserstein距离；两者都可以归结为一个凸优化过程。此外，OT理论揭示了生成器与判别器之间的内在关系是协作的而不是竞争的，并且解释了模式崩溃的根本原因。在此基础上，我们提出了一种新的生成模型，该模型利用自编码器（AE）进行流形学习，并利用OT映射进行概率分布变换。这个AE-OT模型提升了深度学习理论的严谨性和透明性、提高了计算的稳定性和效率，尤其是避免了模式崩溃问题。实验结果验证了我们的假设，并充分展示了我们提出的AE-OT模型的优点。

1. 引言

生成对抗网络（GAN）是无条件图像生成的主要方法之一。在对数据集进行训练后，GAN能够生成逼真的、视觉上吸引人的样本。GAN方法训练了一种无条件生成器和一种判别器，其中生成器可以将随机噪声转换成真实图像，而判别器用于测量生成样本与真实图像之间的差异。GAN已经过多次改进。其中一个突破是将最优传输（OT）理论与GAN相结合，如Wasserstein GAN（WGAN）[1]。在WGAN框架中，生成器计算了从白噪声到数据分布的OT映射，而判别器计算了真实数据分布与生成数据分布之间的Wasserstein距离。

1.1. 流形分布假设

GAN的成功可以通过以下事实进行解释，即GAN有效地发现了真实数据集的内在结构。该结构可以用流形分布假设来表示，即一类特定的自然数据主要集中在一个低维流形上，且该低维流形被嵌入高维背景空间[2]。

图1显示了MNIST数据集的流形结构。每个手写数

* Corresponding author.
E-mail address: gu@cs.stonybrook.edu (X. Gu).
# These authors contributed equally to this work.

英文原文: Engineering 2020, 6(3): 361–374
引用本文: Na Lei, Dongsheng An, Yang Guo, Kehua Su, Shixia Liu, Zhongxuan Luo, Shing-Tung Yau, Xianfeng Gu. A Geometric Understanding of Deep Learning. *Engineering*, https://doi.org/10.1016/j.eng.2019.09.010

字图像的维数为28 × 28，且被看作是$\mathbb{R}^{784}$图像空间中的一个点。MNIST数据集主要集中在一个低维流形（2D流形）附近。通过利用t-SNE流形嵌入算法[3]，MNIST数据集可被映射到一个平面区域上，而且每个图像可被映射到一个点上。表示相同数字的图像被映射到同一个集群中，这里共有10个集群，每个集群分别用不同的颜色编码。这表明MNIST数据集分布在一个二维（2D）曲面附近，该曲面被嵌入在$\mathbb{R}^{784}$的单位超立方体中。

1.2. GAN 理论模型

图2显示了GAN的理论模型。真实数据分布v主要集中在被嵌入背景空间χ中的流形Σ上。(Σ, v) 共同揭示了真实数据集的内在结构。GAN模型计算了隐空间Z到流形Σ的解码映射g_θ，其中，θ表示深度神经网络（DNN）参数。ζ是隐空间中的Gaussian分布，g_θ将ζ前推为μ_θ。判别器计算了真实数据分布v和生成数据分布μ_θ之间的距离，如Wasserstein距离$W_c(\mu_\theta, v)$，其等价于Kontarovich势能φ_ξ。

虽然GAN有很多优点，但是它们也有一些严重的缺点。从理论上讲，我们对深度学习的基本原理的理解仍然比较粗浅。从实践来看，GAN的训练是复杂的，且其对超参数非常敏感，而且GAN经常会遇到模式崩溃问题。最近，Meschede等[4]研究了9种不同的GAN模型及其变体，结果表明，基于梯度下降的GAN优化并不总是局部收敛的。

根据流形分布假设，自然数据集可以被表示为关于流形的概率分布。因此，GAN主要完成两项任务：①流形学习，即计算隐空间与背景空间之间的解码映射和编码映射；②概率变换，即在隐空间或图像空间中计算白噪声与数据分布之间的变换。

图3显示了生成器映射$g_\theta = h \circ T$的分解，其中，h: $Z \to \Sigma$是从隐空间到背景空间中数据流形Σ的解码映射，T: $Z \to Z$是概率分布变换映射。流形学习的解码映射是h，测度变换映射是T。

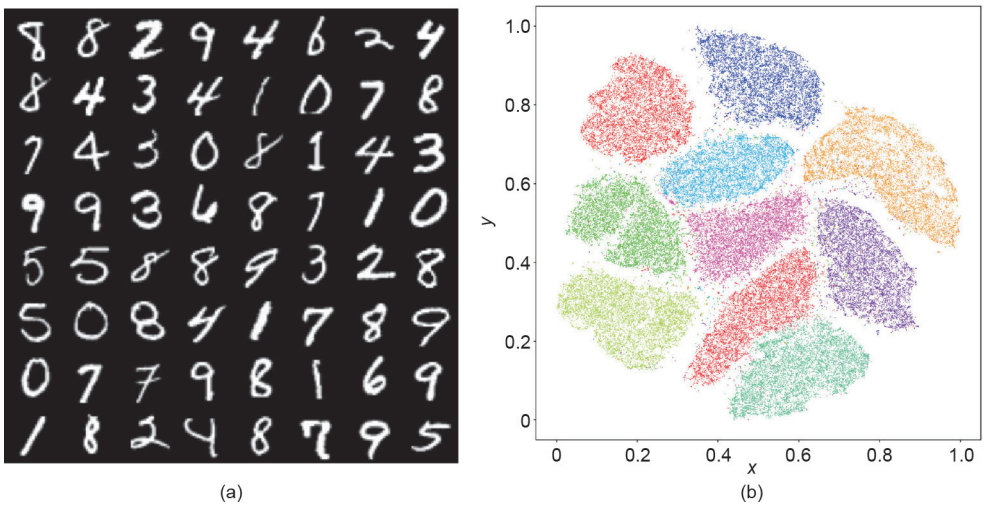
1.3. OT 观点

OT理论[5]研究的是以最经济的方式将一个概率分布转化为另一个概率分布的问题。OT理论给出了计算最优映射的严格而强大的方法，这些方法可以将一个概率分布转换为另一个概率分布，同时计算出它们之间的距离[6]。

如前所述，GAN完成了流形学习和概率分布变换两大任务。后一项任务可以通过直接使用OT方法完成。具体来说，在图3中，概率分布变换映射T可以通过OT理论来计算。判别器计算了真实数据分布和生成数据分布之间的Wasserstein 距离$W_c(\mu_\theta, v)$，这个可以利用OT方法直接计算得到。

从理论角度来看，GAN可以由OT理论来解释，从而使得一部分黑匣子变得透明，同时将概率分布变换过程简化为一个凸优化过程。OT理论使解的存在性和唯一性具有理论保证，而且其收敛速度和近似程度也可以得到全面分析。

OT理论也解释了模式崩溃的根本原因。根据Monge-Ampère方程的正则性理论，变换映射在某些奇



(a) (b)

图1. MNIST数据集的流形分布。（a）MNIST数据集中的手写数字；（b）利用t-SNE算法得到的2D平面内数字的嵌入结果。将x和y相对坐标进行标准化。

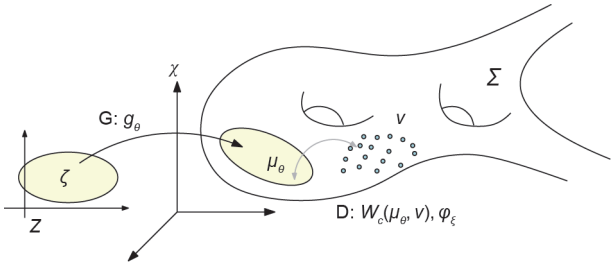


图2. GAN的理论模型。G：生成器；D：判别器。

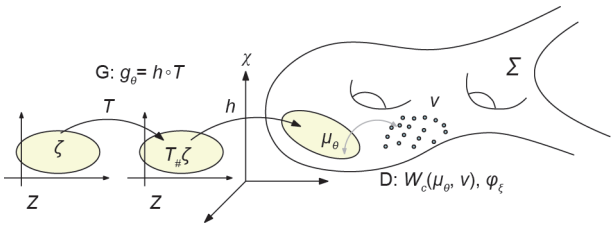


图3. 生成器映射被分解为解码映射h和概率分布变换映射T。$T_\#\zeta$是由T推导出的前推测度。

异集上是不连续的。然而，DNN只能表达连续函数和连续映射。因此，目标变换映射位于GAN所表示的函数空间之外。这种内在的冲突使得模式崩溃问题不可避免。

OT解释还揭示了更复杂的生成器和判别器之间的关系。在现有的GAN模型中，生成器和判别器之间是相互竞争的，它们不共享中间的计算结果。OT理论表明，在L^2成本函数下，生成器和判别器的最优解可以用闭合式来相互表示。因此，生成器与判别器之间的关系应该是相互协作的而不是相互竞争的，而且它们应该共享中间的计算结果以提高计算效率。

1.4. AE-OT 模型

为了降低GAN的训练难度，特别是避免模式崩溃问题，我们提出了一种基于OT理论的更简单的生成模型——自编码（AE）OT模型（AE-OT），如图4所示。

如前所述，生成模型的两个主要任务是流形学习和概率分布变换。AE计算了编码映射$f_\theta : Z \to \Sigma$和解码映射$g_\xi : \Sigma \to Z$, 目的是为了流形学习。OT映射$T : Z \to Z$, 将白噪声ζ变换为由编码映射$(f_\theta)_\#\nu$前推的数据分布。

AE-OT模型有很多优点。从理论上讲，OT理论已经建立并得到了人们的充分理解。通过解耦解码映射和OT映射，我们可以提高生成模型的理论严谨性，从而使部分黑匣子透明化。实际上，OT映射可被简化成一个凸优化问题，从而保证解的存在性和唯一性，同时使

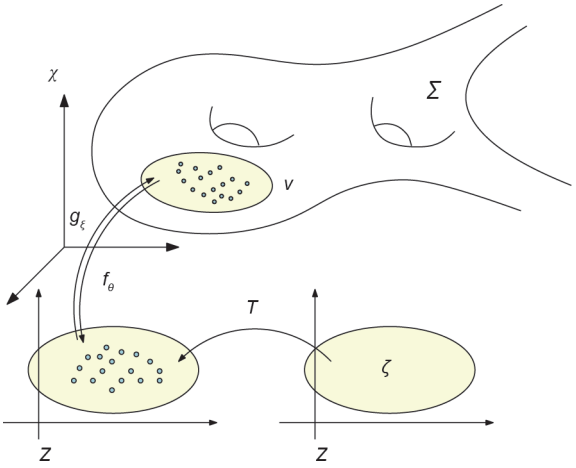


图4. 生成模型AE-OT，将AE和OT相结合。

得训练过程不会仅局限于局部最优；与OT映射相关的凸能量具有明显的Hessian矩阵结构，因此，我们可以利用二阶收敛的牛顿法或超线性收敛的拟牛顿法进行优化。相比之下，现有的生成模型是以具有线性收敛性的梯度下降法为基础的。而且在AE-OT模型中，未知数的个数与训练样本的个数相等，从而避免了过度参数化的问题。在Monte Carlo方法中，采样密度可以完全控制OT映射的误差范围。自适应等级分层算法进一步提高了计算效率。利用图形处理器（GPU）可实现并行OT映射算法。更重要的是，AE-OT模型可以消除模式崩溃问题。

1.5. 贡献

本研究运用OT理论对GAN模型进行了解释。GAN可以完成流形学习和概率分布变换两大任务，后一项任务可以通过OT方法来实现。生成器计算了OT映射，而判别器计算了真实数据分布和生成数据分布之间的Wasserstein距离。使用Brenier定理，我们可以将生成器和判别器之间的竞争关系用协作关系来代替；根据Monge-Ampère方程的正则性理论，分布变换映射的不连续性导致了模式崩溃。我们进一步提出，利用AE-OT模型来解耦流形学习和概率分布变换，从而使部分黑匣子透明化、提高训练效率以及避免模式崩溃。实验结果表明了我们所提出的方法的有效性。

本文的组织结构如下：第2部分简要回顾了OT与GAN的相关工作；第3部分简要介绍了OT的基本理论以及Monge-Ampère方程的正则性理论；第4部分介绍了一种适合深度学习设置的用于计算OT的变分框架；第5部

分从OT的角度分析了GAN模型，解释了生成器与判别器之间的协作关系（不是竞争关系），以及揭示了模式崩溃的内在原因；第6部分总结了实验结果；第7部分对全文进行了总结。

2. 前期工作

2.1. 最优传输

OT问题在各个领域都发挥着重要的作用。详细描述，请读者参照参考文献[7]和[8]。

当输入域和输出域均为Dirac分布时，OT问题可被看作是一种标准线性规划（LP）任务。为了将问题扩展到大数据集，参考文献[9]的作者在原LP问题中增加了一个熵正则化器，则正则化解可以通过Sinkhorn算法被快速计算出来。后来Solomon等[10]通过引入快速卷积提高了计算效率。

第二种解决OT问题的方法是通过OT问题与凸几何之间的联系来最小化凸能量[6]，从而计算出连续测度与逐点测度之间的OT映射。在参考文献[11]中，作者利用Legendre对偶理论将凸几何OT问题与Kantorovich对偶问题联系起来。本文所提出的方法是该方法在高维空间上的一种扩展。如果输入和输出都是连续密度，求解OT问题就等价于求解著名的Monge-Ampère方程，该方程是一个高度非线性椭圆偏微分方程（PDE）。有了一个额外的虚拟时间维度，这个问题可以通过计算流体动力学来解决[12–14]。

2.2. 生成模型

在机器学习领域，能够生成复杂且高维的数据的生成模型近年来变得越来越重要。具体来说，生成模型主要被用于从给定的图像数据集中生成新的图像。在早期研究中，一些方法已被采用，如深度信念网络[15]和深度玻尔兹曼机[16]。然而，这些方法的相关训练通常比较困难和低效。后来，变分AE（VAE）方法取得了重要突破[17]，其中解码器利用变分方法将Gaussian分布逼近了真实数据分布[17,18]。在此基础上，研究人员进行了一系列新的研究工作，包括对偶自编码器（AAE）[19]和Wasserstein AE（WAE）[20]。尽管VAE训练相对容易，但它们生成的图像看起来很模糊。在某种程度上，这是由于显式表达的密度函数可能无法表示真实数据分布的复杂性和无法学习高维数据分布[21,22]。后来，研究人员提出了其他非对抗性训练方法，如PixelCNN [23]、PixelRNN [24]和WaveNet [25]。然而，由于这些方法的自回归性质，新样本的生成是不能并行的。

2.3. 对抗生成模型

针对上述模型的不足，研究人员提出了GAN [26]。虽然GAN是生成逼真样本的强大工具，但是它们很难被训练，而且会出现模式崩溃的问题。为了更好地训练GAN，研究人员已经提出了各种改进措施，包括改变损失函数（如WGAN [1]）以及通过剪切[1]、梯度正则化[4,27]或者光谱归一化[28]来将判别器正则化。然而，GAN的训练仍然是棘手的，需要仔细选择超参数。

2.4. 生成模型的评估

生成模型的评估仍然具有挑战性。早期的工作包括概率标准[29]。然而，最近的生成模型（尤其是GAN）不适合这种评估。传统上，GAN的评估依赖于对少数示例或用户研究的可视化检查。近年来，研究人员提出了几种定量评价标准。Inception score（IS）[30]可同时测量多样性和图像质量，然而它不是距离指标。为了克服IS的缺点，研究人员在参考文献[31]中引入了Fréchet inception distance（FID）。该方法对图像的破坏具有较强的鲁棒性，而且与视觉保真度有很好的相关性。最近的研究[32]介绍了分布的精度和召回率（PRD），这两个指标用于测量真实数据分布和生成数据分布之间的精度和查全率。为了公平地评测GAN，研究人员在参考文献[33]中进行了大规模比较，在统一的网络架构下，研究人员比较了7种不同的GAN和VAE，并建立了一个通用的评价标准。

2.5. 非对抗性方法

最近，研究人员也提出了各种非对抗性的方法。生成潜优化（GLO）[34]是一种“无编码器AE”的方法，其中生成模型通过非对抗性损失函数进行训练，并且取得了比VAE更好的结果。隐式最大似然估计（IMLE）[35]是一种最近点迭代（ICP）相关的生成模型训练方法。后来Hoshen和Malik [36]提出了生成式隐含最近邻（GLANN），该方法结合了GLO和GLANN的优点。该方法首先利用GLO发现了从图像空间到隐空间的嵌入，然后利用IMLE计算出了任意分布与隐藏代码之间的转换。

其他一些方法则是利用含有可控Jacobian矩阵的DNN直接逼近了从噪声空间到图像空间的分布变换映射[37–39]。近年来，研究人员选择了一些基于能量的模型[40–42]，他们利用DNN来表示能量函数，并通过Gibb分布对图像分布进行建模。这些方法利用现有模型交替生成伪样本，然后利用生成的伪样本和真实样本对模型参数进行优化。

3. OT 理论

在本章中，我们将介绍经典OT理论中的基本概念和定理，重点介绍Brenier方法及其在离散集中的推广。具体细节可参考Villani的专著[5]。

3.1. Monge 问题

假设$X \subset \mathbb{R}^d$, $Y \subset \mathbb{R}^d$是两个d维Euclidean空间$\mathbb{R}^d$的子集，μ和υ是被分别定义在X和Y上的两个概率测度，则密度函数如下：

$\mu(\boldsymbol{x}) = f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$

$v(\boldsymbol{y}) = g(\boldsymbol{y})\mathrm{d}\boldsymbol{y}$

假设总测度相等，即$\mu(X) = \upsilon(Y)$，那么

$$\int_X f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_Y g(\boldsymbol{y})\mathrm{d}\boldsymbol{y} \tag{1}$$

我们只考虑保测度的映射。

Defnition 3.1（保测度映射）。如果对于任何可测集$B \subset Y$，集合$T^{-1}(B)$是μ-可测的，并且$\mu[T^{-1}(B)] = \upsilon(B)$，那么映射$T: X \to Y$是保测度的，即

$$\int_{T^{-1}(B)} f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_B g(\boldsymbol{y})\mathrm{d}\boldsymbol{y} \tag{2}$$

保测度条件被记作$T_{\#}\mu = \upsilon$，其中$T_{\#}\mu$为T诱导的前推测度。

给定成本函数$c(\boldsymbol{x}, \boldsymbol{y}): X \times Y \to \mathbb{R}_{\geqslant 0}$，该函数表示从源到目标的传输每个单位质量的代价，则定义映射$T: X \to Y$的总传输代价为

$$C_{\mathrm{t}} = \int_X c[\boldsymbol{x}, T(\boldsymbol{x})]\mathrm{d}\mu(\boldsymbol{x}) \tag{3}$$

Monge的OT问题在于寻找使总传输成本最小的保测度映射。

Problem 3.2（Monge's [43]；MP）。给定传输成本函数$c(\boldsymbol{x}, \boldsymbol{y}): X \times Y \to \mathbb{R}_{\geqslant 0}$，求使总传输成本最小的保测度映射$T: X \to Y$，即

$$(\mathrm{MP}) \min_{T_{\#}\mu = \upsilon} \int_X c[\boldsymbol{x}, T(\boldsymbol{x})]\mathrm{d}\mu(\boldsymbol{x}) \tag{4}$$

Defnition 3.3（OT映射）。Monge的问题的解被称为OT映射。OT映射的总传输成本被称为μ和υ之间的Wasserstein距离，被记作$W_c(\mu, \upsilon)$。

$$W_c(\mu, \upsilon) = \min_{T_{\#}\mu = \upsilon} \int_X c[\boldsymbol{x}, T(\boldsymbol{x})]\mathrm{d}\mu(\boldsymbol{x}) \tag{5}$$

3.2. Kontarovich 的方法

根据成本函数及其测度的性质，(X, μ)和(Y, υ)之间的OT映射可能不存在。Kontarovich将传输映射扩展至传输平面，并定义了联合概率测度$\rho(\boldsymbol{x}, \boldsymbol{y}): X \times Y \to \mathbb{R}_{\geqslant 0}$，这样$\rho$的边际概率分别等于$\mu$和$\upsilon$。令投影映射$\pi_{\boldsymbol{x}}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}$和$\pi_{\boldsymbol{y}}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{y}$，然后定义联合测度类如下：

$$\begin{aligned}\Pi(\mu, \upsilon) = \Big\{\rho(\boldsymbol{x}, \boldsymbol{y}) : X \times Y \to \mathbb{R} : (\pi_{\boldsymbol{x}})_{\#}\rho = \\ \mu, (\pi_{\boldsymbol{y}})_{\#}\rho = \upsilon\Big\}\end{aligned} \tag{6}$$

Problem 3.4（Kontarovich；KP）。给定一个传输成本函数$c(\boldsymbol{x}, \boldsymbol{y}): X \times Y \to \mathbb{R}_{\geqslant 0}$，求得联合概率测度$\rho(\boldsymbol{x}, \boldsymbol{y}): X \times Y \to \mathbb{R}_{\geqslant 0}$，使得传输总成本最小。

$$(\mathrm{KP})\ W_c(\mu, \upsilon) = \min_{\rho \in \Pi(\mu, \upsilon)} \int_{X \times Y} c(\boldsymbol{x}, \boldsymbol{y})\mathrm{d}\rho(\boldsymbol{x}, \boldsymbol{y}) \tag{7}$$

Kontarovich的问题（KP）可以采用LP方法来求解。由于LP的对偶性，方程（7）（KP公式）可以被重新表述为对偶问题（DP），具体如下：

Problem 3.5（对偶；DP）。给定一个传输成本函数$c(\boldsymbol{x}, \boldsymbol{y}): X \times Y \to \mathbb{R}_{\geqslant 0}$，求得真实函数$\varphi: X \to \mathbb{R}$和$\psi: Y \to \mathbb{R}$，使得

$$\begin{aligned}(\mathrm{DP})\ \max_{\varphi, \psi}\Big[\int_X \varphi(\boldsymbol{x})\mathrm{d}\mu + \int_Y \psi(\boldsymbol{y})\mathrm{d}v : \\ \varphi(\boldsymbol{x}) + \psi(\boldsymbol{y}) \leqslant c(\boldsymbol{x}, \boldsymbol{y})\Big]\end{aligned} \tag{8}$$

公式（8）的最大值给出了Wasserstein距离。现有的WGAN模型大多是基于L^1成本函数下的对偶形式。

Defnition 3.6（c-变换）。$\varphi : X\to\mathbb{R}$的c-变换被定义为$\varphi^c : Y\to\mathbb{R}$：

$$\varphi^c(\boldsymbol{y}) = \inf_{x\in X}[c(\boldsymbol{x},\boldsymbol{y}) - \varphi(\boldsymbol{x})] \tag{9}$$

则对偶问题可被重新表述为如下形式：

$$\text{(DP) } W_c(\mu,\nu) = \max_{\varphi}\int_X \varphi(\boldsymbol{x})\mathrm{d}\mu + \int_Y \varphi^c(\boldsymbol{y})\mathrm{d}\nu \tag{10}$$

3.3. Brenier 的方法

对于二次Euclidean距离成本函数，Brenier [44]证明了OT映射的存在性、唯一性和内在结构。

Theorem 3.7（Brenier's [44]）。假设X和Y是Euclidean空间$\mathbb{R}^d$中的子集，并且传输成本是Euclidean 距离的平方，即$c(\boldsymbol{x},\boldsymbol{y}) = 1/2||\boldsymbol{x}-\boldsymbol{y}||^2$。此外，$\mu$是绝对连续的，而且$\mu$和$\upsilon$存在有限的二阶矩

$$\int_X \|\boldsymbol{x}\|^2\mathrm{d}\mu(\boldsymbol{x}) + \int_Y \|\boldsymbol{y}\|^2\mathrm{d}\nu(\boldsymbol{y}) < \infty \tag{11}$$

则存在一个凸函数u: $X\to\mathbb{R}$，即所谓的Briener势能，其梯度映射∇u给出了Monge问题的解：

$$(\nabla u)_{\#}\mu = \nu \tag{12}$$

由于Brenier势能在常数范围内是唯一的，因此OT映射是唯一的。

假设Brenier势能是C^2光滑的，则它是下面Monge-Ampère方程的解。

$$\det\left[\frac{\partial^2 u(\boldsymbol{x})}{\partial\boldsymbol{x}_i\partial\boldsymbol{x}_j}\right] = \frac{f(\boldsymbol{x})}{g\mathring{\nabla}u(\boldsymbol{x})} \tag{13}$$

在$\mathbb{R}^d$中，对于Euclidean空间上的L^2传输成本函数$c(\boldsymbol{x},\boldsymbol{y}) = 1/2||\boldsymbol{x}-\boldsymbol{y}||^2$，$c$-变换与经典Legendre变换之间有着特殊的关系。

Defnition 3.8（Legendre变换）。给定一个函数φ: $\mathbb{R}^n\to\mathbb{R}$，其Legendre变换被定义为如下形式：

$$\varphi^*(\boldsymbol{y}) = \sup_{\boldsymbol{x}}[\langle\boldsymbol{x},\boldsymbol{y}\rangle - \varphi(\boldsymbol{x})] \tag{14}$$

由此可知，当$c(\boldsymbol{x},\boldsymbol{y}) = 1/2||\boldsymbol{x}-\boldsymbol{y}||^2$时，下面的等式成立。

$$\frac{1}{2}\|\boldsymbol{y}\|^2 - \varphi^c(\boldsymbol{y}) = \left[\frac{1}{2}\|\boldsymbol{x}\|^2 - \varphi(\boldsymbol{x})\right]^* \tag{15}$$

Theorem 3.9（Brenier极分解[44]）。假设X和Y是Euclidean空间$\mathbb{R}^d$，μ相对于Lebesgue测度是绝对连续的，且映射φ: $X\to Y$将μ前推为ν，即$\varphi_{\#}\mu = \upsilon$，则存在一个凸函数$u$: $X\to\mathbb{R}$，使得$\varphi = \nabla u\circ s$。式中，$s$: $X\to X$是保测度的，即$s_{\#}\mu = \mu$。此外，这个分解是唯一的。

下面的定理在OT理论中是众所周知的。

Theorem 3.10（Villani [5]）。给定凸紧区域$\Omega\subset\mathbb{R}^d$上定义的测度$\mu$和$\upsilon$，这里存在一个成本函数为$c(\boldsymbol{x},\boldsymbol{y}) = h(\boldsymbol{x}-\boldsymbol{y})$的OT平面$\rho$，其中$h$是严格凸的。假定$\mu$是绝对连续的，并且$\partial\Omega$为零测度，则$\rho$是唯一的，且其具有$(\mathrm{id}, T_{\#})\mu$（id：恒等映射）的形式。另外，这里存在一个Kantorovich势能φ，而且映射T可用下式表示为：

$$T(\boldsymbol{x}) = \boldsymbol{x} - (\nabla\boldsymbol{h})^{-1}[\nabla\varphi(\boldsymbol{x})]$$

当$c(\boldsymbol{x},\boldsymbol{y}) = 1/2||\boldsymbol{x}-\boldsymbol{y}||^2$时，我们可以得到

$$T(\boldsymbol{x}) = \boldsymbol{x} - \nabla\varphi(\boldsymbol{x}) = \nabla\left[\frac{1}{2}\|\boldsymbol{x}\|^2 - \varphi(\boldsymbol{x})\right] = \nabla u(\boldsymbol{x})$$

在这种情况下，Brenier势能u和Kantorovich势能φ有如下关系：

$$u(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|^2 - \varphi(\boldsymbol{x}) \tag{16}$$

3.4. OT 映射的正则性

令Ω和Λ是$\mathbb{R}^d$中两个有边界的光滑开集，令$\mu = f\mathrm{d}\boldsymbol{x}$和$\nu = g\mathrm{d}\boldsymbol{y}$是$\mathbb{R}^d$上两个概率测度，那么$f|_{\mathbb{R}^d\Omega} = 0$和$g|_{\mathbb{R}^d\Lambda} = 0$。设$f$和$g$在$\Omega$和$\Lambda$上分别是非零和非无穷的。

3.4.1. 凸目标域

Defnition 3.11（Hölder 连续）。一个实值函数或复值函数f在d维Euclidean空间中满足Hölder条件，或者它是Hölder连续时，此时存在非负实常数C，且$\alpha > 0$，使得$|f(\boldsymbol{x}) - f(\boldsymbol{y})|\leqslant C||\boldsymbol{x}-\boldsymbol{y}||^{\alpha}$对于$f$定义域中的所有$x$和$y$都成立。

Defnition 3.12（Hölder空间）。Hölder空间为$C^{k,\alpha}(\Omega)$，其中Ω是某个Euclidean空间的一个开子集，并且整数$k\geqslant 0$，它是由在Ω上有直到k阶连续偏导数的函数组成，从而使得k阶偏导数是α阶Hölder连续的，且$0 < \alpha \leqslant 1$。$C^{k,\alpha}(\Omega)$意味着上述条件适用于Ω的任意紧子集。

Theorem 3.13（Caffarelli [45]）。如果Λ是凸的，那么Brenier势能u是严格凸的，此外，

（1）如果$\lambda\leqslant f$，当$\lambda > 0$时，$g\leqslant 1/\lambda$，那么$u\in C^{k,\alpha}_{\mathrm{loc}}(\Omega)$。

（2）如果$f\in C^{k,\alpha}_{\mathrm{loc}}(\Omega)$和$g\in C^{k,\alpha}_{\mathrm{loc}}(\Lambda)$，并且$f, g > 0$，那么$u\in C^{k+2,\alpha}_{\mathrm{loc}}(\Omega)$且$[k\geqslant 0, a\in(0.1)]$。

3.4.2. 非凸目标域

如果Λ是非凸的且存在光滑的f和g，那么$u \notin C^1(\Omega)$，而且OT映射∇u在奇异点处是非连续的。

Defnition 3.14（次梯度）。给定开区间$\Omega \subset \mathbb{R}^d$和一个凸函数$u: X \to \mathbb{R}$，对于$\boldsymbol{x} \in \Omega$，$u$在$\boldsymbol{x}$点的次梯度（次微分）可被定义为如下形式：

$$\partial u(\boldsymbol{x}) = \{\boldsymbol{p} \in \mathbb{R}^n : u(\boldsymbol{z}) \geq u(\boldsymbol{x}) + \langle \boldsymbol{p}, \boldsymbol{z} - \boldsymbol{x} \rangle, \ \forall \boldsymbol{z} \in \boldsymbol{\Omega}\}$$

显然，$u(x)$是一个闭凸集。从几何学来看，如果$p \in u(\boldsymbol{x})$，那么超平面$l_{x,\boldsymbol{p}}(\boldsymbol{z}) = u(\boldsymbol{x}) + \langle \boldsymbol{p}, \boldsymbol{z} - \boldsymbol{x} \rangle$在$\boldsymbol{x}$点从下方触碰到了$u$，即$\Omega$中的$l_{x,\boldsymbol{p}} \leqslant u$并且$l_{x,\boldsymbol{p}}(\boldsymbol{x}) = u(\boldsymbol{x})$，其中$l_{x,\boldsymbol{p}}$是$u$在$\boldsymbol{x}$点处的支撑平面。

如果Brenier势能u的次梯度$\partial u(\boldsymbol{x})$包含一个点，则$u$在$\boldsymbol{x}$点处可微。我们根据次梯度的维数对这些点进行分类，并且定义集合$\Sigma_k(u) = \left\{\boldsymbol{x} \in \mathbb{R}^d | \dim[\partial u(\boldsymbol{x})] = k\right\}$，$k = 0, 1, 2, ..., d$。

可以看出，$\Sigma_0(u)$是正则点的集合，而$\Sigma_k(u)$是奇异点的集合，其中$k > 0$。我们也定义了$\boldsymbol{x}$点的可达次梯度，具体如下：

$$\nabla_* u(\boldsymbol{x}) = \left\{\lim_{k \to \infty} \nabla u(\boldsymbol{x}_k) | \boldsymbol{x}_k \in \Sigma_0, \boldsymbol{x}_k \to \boldsymbol{x}\right\}$$

由此可知，次梯度等于可达次梯度的凸包，即

$$\partial u(\boldsymbol{x}) = \text{Convex hull } \nabla_* u(\boldsymbol{x})$$

Theorem 3.15（正则性）。令$\Omega, \Lambda \subset \mathbb{R}^d$为两个有边界的开集，并且令$f, g: \mathbb{R}^d \to \mathbb{R}^+$为两个概率密度函数，该密度函数在$\Omega$和$\Lambda$之外为0，而在$\Omega$和$\Lambda$上则界于0和无穷之间。Theorem 3.7中的OT映射被表示为$T = \nabla u: \Omega \to \Lambda$。那么存在两个相对闭集$\Sigma_\Omega \subset \Omega$和$\Sigma_\Lambda \subset \Lambda$，且$\Sigma_\Omega = \Sigma_\Lambda = 0$，当常数$\alpha > 0$时，使得$T: \Omega\Sigma_\Omega \to \Lambda\Sigma_\Lambda$是属于$C^{0,\alpha}_{\text{loc}}$类的拓扑同胚。

我们称Σ_Ω为OT映射$\nabla u: \Omega \to \Lambda$的奇异集。图5给出了基于Theorem 4.2的算法所计算出的奇异点集结构。具体形式如下：

$$\Sigma_0 = \boldsymbol{\Omega}\{\{\Sigma_1 \cup \Sigma_2\}, \ \Sigma_1 = \bigcup_{k=0}^{3} \gamma_k, \ \Sigma_2 = \{\boldsymbol{x}_0, \boldsymbol{x}_1\}$$

$\boldsymbol{x}_0$点的次梯度$\partial u(\boldsymbol{x}_0)$整个覆盖了$\Lambda$内部孔洞，而$\partial u(\boldsymbol{x}_1)$覆盖了阴影三角形区域。对于$\gamma_k(t)$上的每个点，$\partial u[\gamma_k(t)]$是$\Lambda$外部的一条线段。$\boldsymbol{x}_1$是$\gamma_1$、$\gamma_2$和$\gamma_3$的分歧点。Brenier势能在$\Sigma_1$和$\Sigma_2$上是不可微的，OT映射$\nabla u$在$\Sigma_1$和$\Sigma_2$上是不连续的。

4. 计算方法

Brenier定理可以被直接推广到离散情形中。在GAN模型中，源测度μ是一个被定义在紧凸集Ω上的均匀（或高斯）分布；目标测度v被表示为经验测度，它是Dirac测度的总和，即

$$v = \sum_{i=1}^{n} v_i \delta(\boldsymbol{y} - \boldsymbol{y}_i) \tag{17}$$

式中，$Y = \{\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_n\}$是训练样本，其权重为$\Sigma_{i=1}^n v_i = \mu(\boldsymbol{\Omega})$；$\delta$是特征函数。

每个训练样本$\boldsymbol{y}_i$对应一个Brenier势能的支撑平面，且用下式表示，即

$$\pi_{\boldsymbol{h},i}(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{y}_i \rangle + \boldsymbol{h}_i \tag{18}$$

式中，支撑平面的截距（高度）$\boldsymbol{h}_i$是未知变量。我们将所有的高度变量记为$\boldsymbol{h} = (\boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_n)$。

Euclidean空间中一族超平面的包络是一个超曲面，它与该族的每个成员都相切于某一点，这些切点共同构成了整个包络超曲面。如图6所示，Brenier势能$u_{\boldsymbol{h}}: \Omega \to \mathbb{R}$是一个由$\boldsymbol{h}$确定的分片线性凸函数，这个凸函数是它所有支撑平面的上包络，即

$$u_{\boldsymbol{h}}(\boldsymbol{x}) = \max_{i=1}^{n} [\pi_{\boldsymbol{h},i}(\boldsymbol{x})] = \max_{i=1}^{n} [\langle \boldsymbol{x}, \boldsymbol{y}_i \rangle + \boldsymbol{h}_i] \tag{19}$$

Brenier势能图是一个凸多面体。每一个支撑平面$\pi_{\boldsymbol{h},i}$对应多面体的一个面。多面体的投影诱导了Ω的一个单元分解，其中每个支撑平面$\pi_i(\boldsymbol{x})$的投影形成一个单元$W_i(\boldsymbol{h})$，而$\boldsymbol{p}$是$\mathbb{R}^d$中的任意一点，具体如下：

$$\boldsymbol{\Omega} = \bigcup_{i=1}^{n} W_i(\boldsymbol{h}) \cap \boldsymbol{\Omega}, \ W_i(\boldsymbol{h}) = \{\boldsymbol{p} \in \mathbb{R}^d | \nabla u_{\boldsymbol{h}}(\boldsymbol{p}) = \boldsymbol{y}_i\} \tag{20}$$

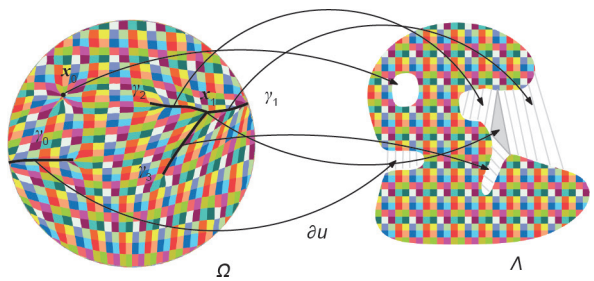


图5. OT映射的奇异点集结构。

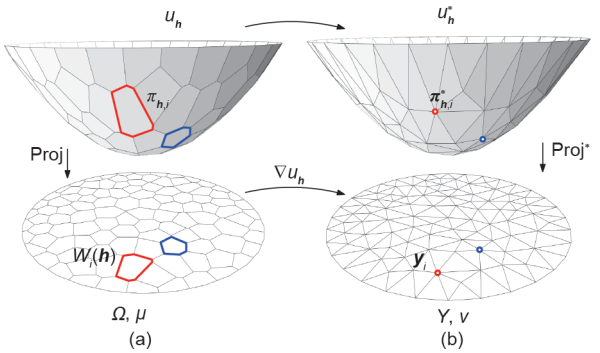


图6. 分片线性Brenier势能函数（a）及其Legendre变换u_h^*（b）。$\boldsymbol{\delta}_{h,i}^*$: $\pi_{h,i}$的Legendre对偶；∇: u_h的梯度；Proj：投影映射；Proj*：Legendre对偶空间内的投影映射。

这个单元分解是一个功率图。$W_i \cap \Omega$ 的μ测度被记为$w_i(\boldsymbol{h})$，即

$$w_i(\boldsymbol{h}) = \mu[W_i(\boldsymbol{h}) \cap \Omega] = \int_{W_i(\boldsymbol{h})\cap\Omega} \mathrm{d}\mu \tag{21}$$

梯度映射$\nabla u_h : \Omega \to Y$将每个单元$W_i(\boldsymbol{h})$映射为一个点$\boldsymbol{y}_i$，即

$$\nabla u_h : W_i(\boldsymbol{h}) \to \boldsymbol{y}_i,\ i = 1, 2, ..., n. \tag{22}$$

如果公式（17）中目标测度v已知，则由公式（19）可得到一个离散的Brenier势能，且该势能的每个支撑平面$w_i(\boldsymbol{h})$投影的μ-体积等于给定的目标测度v_i。这个结论已被Alexandrov [46]在凸几何中证明。

Theorem 4.1（Alexandrov [46]）。假设Ω是一个紧凸多面体，其在$\mathbb{R}^n$中内部非空；$\boldsymbol{n}_1, \ldots, \boldsymbol{n}_k \subset \mathbb{R}^{n+1}$是$k$个不同的单位向量；第$(n+1)$个坐标是负的以及$v_1, ..., v_k > 0$，使得$\Sigma_{i=1}^k v_i = \mathrm{vol}(\Omega)$。则存在凸多面体$P \subset \mathbb{R}^{n+1}$恰有$k$个余维数为1的平面$F_1, \ldots, F_k$，使得$\boldsymbol{n}_i$是$F_i$的法向量，且$\Omega$与$F_i$投影之间的交集体积为$v_i$。此外，$P$在垂直平移下唯一。

Alexandrov对解的存在的证明是以代数拓扑为基础进行的，其不具构造性。最近，Gu等[6]基于变分方法给出了构造性证明。

Theorem 4.2（参考文献[6]）。令μ是一个被定义在$\mathbb{R}^d$中紧凸区域Ω上的概率测度，令$Y = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n\}$是$\mathbb{R}^d$中的一组不同点。那么，对于任意$v_1, v_2, \ldots, v_n > 0$，其中$\Sigma_{i=1}^n v_i = \mu(\Omega)$，存在$\boldsymbol{h} = (\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_n) \in \mathbb{R}^n$，在相差一个常数$(c, c, \ldots, c)$的意义下唯一，使得对于所有$1 \leqslant i \leqslant n$，$w_i(\boldsymbol{h}) = v_i$。向量$\boldsymbol{h}$是以下凸能量的唯一最小变元，

$$E(\boldsymbol{h}) = \int_0^{\boldsymbol{h}} \sum_{i=1}^n w_i(\eta)\mathrm{d}\eta_i - \sum_{i=1}^n \boldsymbol{h}_i v_i \tag{23}$$

在开凸集上被定义为

$$\boldsymbol{h} = \{\boldsymbol{h} \in \mathbb{R}^n : w_i(\boldsymbol{h}) > 0,\ i = 1, 2, ..., n\} \tag{24}$$

此外，∇u_h在所有的传输映射$T_{\#}\mu = v$中的最小化二次成本为

$$\frac{1}{2}\int_\Omega \| \boldsymbol{x} - T(\boldsymbol{x}) \|^2 d\mu(\boldsymbol{x}) \tag{25}$$

公式（23）中上述凸能量的梯度由下式给出。

$$\nabla E(\boldsymbol{h}) = [w_1(\boldsymbol{h}) - v_1, w_2(\boldsymbol{h}) - v_2, ..., w_n(\boldsymbol{h}) - v_n]^{\mathrm{T}} \tag{26}$$

能量的第i行和第j列的Hessian元素可由下式给出。

$$\frac{\partial w_i}{\partial \boldsymbol{h}_j} = -\frac{\mu(W_i \cap W_j \cap \boldsymbol{\Omega})}{\| \boldsymbol{y}_i - \boldsymbol{y}_j \|},\ \frac{\partial w_i}{\partial \boldsymbol{h}_i} = \sum_{j \neq i} \frac{\partial w_i}{\partial \boldsymbol{h}_j} \tag{27}$$

如图6所示，Hessian矩阵具有明确的几何意义。图6（a）显示了离散的Brenier势能u_h，图6（b）显示了Hessian矩阵由Definition 3.8所定义的Legendre变换u_h^*。Legendre变换可以用几何方法来构造，即对于每个支撑平面$\pi_{h,i}$，我们构造了对偶点$\pi_{h,i}^* = (\boldsymbol{y}_i, \boldsymbol{h}_i)$，其中对偶点的凸包 $\{\pi_{h,1}^*, \pi_{h,2}^*, \ldots, \pi_{h,n}^*\}$是Legendre变换$u_h^*$的图。

u_h^*的投影诱导了$Y = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n\}$的加权Delaunay三角剖分。如图7所示，公式（20）中的power diagram和加权Delaunay三角剖分是彼此的Poincaré对偶，即在power diagram中，如果$W_i(\boldsymbol{h})$和$W_j(\boldsymbol{h})$相交于某个$(d-1)$维单元，则在加权的Delaunay三角剖分中，$\boldsymbol{y}_i$与$\boldsymbol{y}_j$相连。公式（27）中Hessian矩阵的元素是power diagram中$(d-1)$维单元的μ-体积与加权Delaunay三角剖分中对偶边的长度之间的比率。

传统的power diagram与上述定理密切相关。

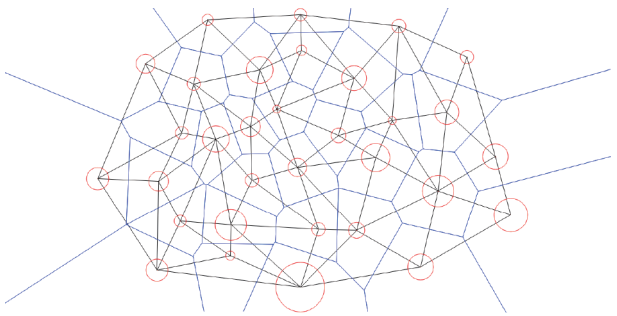


图7. Power diagram（蓝色）和其对偶加权Delaunay三角剖分（黑色）。

Defnition 4.3. (power 距离)。给定具有power权重ψ_i的点$\boldsymbol{y}_i \in \mathbb{R}^n$，power距离可由下式给出。

$$\mathrm{pow}(\boldsymbol{x},\boldsymbol{y}_i) = \| \boldsymbol{x} - \boldsymbol{y}_i \|^2 - \psi_i \tag{28}$$

Defnition 4.4. (power diagram)。给定加权点 $(\boldsymbol{y}_1, \psi_1)$,..., $(\boldsymbol{y}_k, \psi_k)$，power diagram是$\mathbb{R}^d$的单元分解，即

$$\mathbb{R}^d = \bigcup_{i=1}^{k} W_i(\psi) \tag{29}$$

这里的每个单元都是凸多面体，即

$$W_i(\psi) = \{\boldsymbol{x} \in \mathbb{R}^d | \mathrm{pow}(\boldsymbol{x},\boldsymbol{y}_i) \leqslant \mathrm{pow}(\boldsymbol{x},\boldsymbol{y}_j)\} \tag{30}$$

加权Delaunay三角剖分用$T(\psi)$表示，它是power diagram的Poincaré对偶，如果$W_i(\psi) \cap W_j(\psi) \neq \phi$，则在加权Delaunay三角剖分中存在连接$\boldsymbol{y}_i$和$\boldsymbol{y}_j$的边。注意，$pow(\boldsymbol{x}, \boldsymbol{y}_i) \leqslant pow(\boldsymbol{x}, \boldsymbol{y}_j)$等价于

$$\langle \boldsymbol{x},\boldsymbol{y}_i \rangle + \frac{1}{2}\left(\psi_i - \| \boldsymbol{y}_i \|^2\right) \geqslant \langle \boldsymbol{x},\boldsymbol{y}_j \rangle + \frac{1}{2}\left(\psi_j - \| \boldsymbol{y}_j \|^2\right) \tag{31}$$

令$h_i = 1/2(\psi_i - \|\boldsymbol{y}_i\|^2)$，我们将$W_i(\psi)$的定义重写为

$$W_i(\psi) = \{\boldsymbol{x} \in \mathbb{R}^d | \langle \boldsymbol{x},\boldsymbol{y}_i \rangle + \boldsymbol{h}_i \geqslant \langle \boldsymbol{x},\boldsymbol{y}_j \rangle + h_j, \ \forall j\} \tag{32}$$

在实践中，我们的目标是通过优化凸能量方程（23）来计算离散Brenier势能方程（19）。对于低维情况，我们可以通过计算梯度方程（26）和Hessian矩阵方程（27）来直接使用牛顿法。对于深度学习的应用，直接计算Hessian矩阵是不可行的，我们可以使用梯度下降法或超线性收敛的拟牛顿法。梯度下降法的关键是估计μ-体积$w_i(\boldsymbol{h})$。我们可以通过使用Monte-Carlo方法来完成，即我们从分布μ中随机抽取n个样本，并计算落入$W_i(\boldsymbol{h})$的样本数，该样本数是收敛到μ-体积的比率。此方法是完全并行的，并可以通过GPU来实现。此外，我们可以使用等级分层方法来进一步提高效率。首先，我们将目标样本按聚类簇进行分类，然后计算目标样本到聚类簇质心的OT映射；其次，对于每个聚类簇，我们计算了从相应单元到聚类簇内原始目标样本的OT映射。

为了避免模式崩溃，我们需要找到Ω中的奇异点集。如图8所示，目标Dirac测度有两个聚类簇，源是单位平面圆盘上的均匀分布。Brenier势能函数的图是中间带有脊线的凸多面体。脊线在圆盘上的投影是奇异点集$\Sigma_1(u)$，OT映射在Σ_1上是不连续的。在一般情况下，如果两个单元$W_i(\boldsymbol{h})$和$W_j(\boldsymbol{h})$相邻，那么我们可计算相应支撑平面的法线之间的角度为：

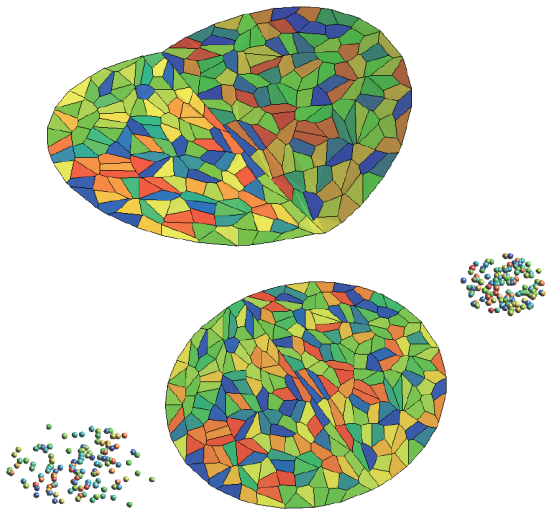


图8. Brenier势能函数的奇异点集与OT映射的间断点集。

$$\theta_{i,j} = \frac{\langle \boldsymbol{y}_i,\boldsymbol{y}_j \rangle}{\| \boldsymbol{y}_i \| \| \boldsymbol{y}_j \|}$$

如果θ_{ij}大于阈值，则公共平面$W_i(\boldsymbol{h}) \cap W_j(\boldsymbol{h})$位于不连续奇点集中。

5. GAN 和最优传输

OT理论为GAN奠定了理论基础。最近的研究成果，如WGAN [1]、WGAN-GP [27]和RW-GAN [47]，都使用了Wasserstein距离来度量已生成的数据分布与实际数据分布之间的偏差。

从OT角度来看，生成器与判别器最优解之间存在一个闭合式，因此生成器与判别器之间应该是相互合作的而不是竞争的。更多细节见参考文献[11]。此外，Monge-Ampère解的正则性理论可以解释GAN的模式崩溃[48]。

5.1. 竞争与合作

图2显示了WGAN [1]的OT视图。根据流形分布假设，真实数据分布υ与一个被嵌入背景空间χ中的流形Σ非常接近。生成器计算了从隐空间Z到背景空间的解码映射g_θ，并且把白噪声ζ（如Gaussian分布）变换为生成分布μ_θ。通过计算Kantorovich势能φ_ξ，判别器计算了μ_θ

和真实分布υ之间的Wasserstein距离$W_c(\mu_\theta, \upsilon)$。$g_\theta$和$\varphi_\xi$都是通过DNN来实现的。

在训练过程中，生成器通过优化g_θ以使$(g\theta)\#\zeta$能更好地逼近υ；判别器通过优化Kantorovich势能φ_ξ来改善对Wassertein距离的估计，生成器和判别器相互竞争、不共享中间结果。在L^1成本函数下，WGAN的交替训练过程可以被看作是期望值的最小-最大优化过程：

$$\min_\theta \max_\xi E_{z\sim\zeta}\{\varphi_\xi[g_\theta(\boldsymbol{z})]\} + E_{y\sim\upsilon}[\varphi_\xi^c(\boldsymbol{y})]$$

但是如果我们把成本函数换成L^2距离，那么根据Theorem 3.10，在最优情况下，Briener势能u和Kontarovic 势能能能φ是通过公式（16）的闭合式$u(\boldsymbol{x}) = 1/2||\boldsymbol{x}||^2 - \varphi(\boldsymbol{x})$相联系的。生成器寻找到了OT映射$\nabla u$，而判别器计算出了$\varphi$。因此，一旦生成器达到最优解，判别器无需任何训练即可得到最优解，反之亦然。

更详细地说，假设在第k次迭代中，生成器映射为g_θ^k。判别器计算了Kontarovich势能φ_ξ，其给出了当前生成的数据分布$(g_\theta^k)_\#\zeta$与实数据分布υ之间的Wasserstein距离；∇u给出了从$(g_\theta^k)_\#\zeta$到υ的OT映射。因此我们可以得到：

$$\upsilon = (\nabla u)_\#\left[(g_\theta^k)_\#\zeta\right] = (\nabla u \mathring{g}_\theta^k)_\#\zeta = \left[\left(\mathrm{id} - \nabla_{\varphi_\xi}\right)\mathring{g}_\theta^k\right]_\#\zeta$$

这意味着生成器映射可以被更新为

$$g_\theta^{k+1} = (\mathrm{id} - \nabla\varphi_\xi)\mathring{g}_\theta^k \qquad (33)$$

这个结论表明，原则上我们可以跳过生成器的训练过程；在实际应用中，我们通过共享中间计算结果可以大大提高计算效率。因此，在设计GAN架构时，协作优于竞争。

5.2. 模式崩溃和正则性

尽管GAN在许多应用中十分强大，但是它们有十分致命的缺陷。第一，GAN的训练比较复杂，其对超参数敏感以及收敛性差；第二，GAN易产生模式崩溃问题；第三，GAN可能会产生不真实的样本。不收敛性差、模式崩溃和生成不真实的样本等问题都可以通过OT映射的正则性定理来解释。

根据Brenier的极分解定理，即Theorem 3.9，任何保测度映射都可以被分解为两个映射，其中一个是OT映射，它是Monge-Ampère方程的解。根据正则性Theorem 3.15，如果目标测度υ的支集Λ具有多个连通分支，即υ具有多个模式，或者Λ是非凸集合，那么OT映射T: $\Omega\to\Lambda$在奇异点集Σ_Ω上是不连续的。

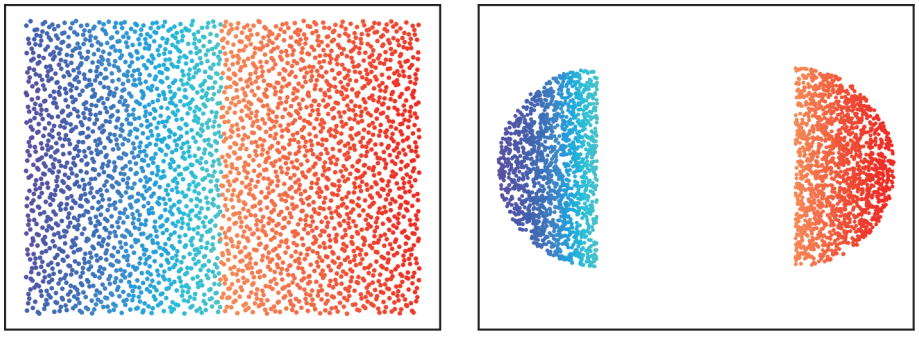
图9显示了多个连通的情形，Λ具有两个连通分支，OT映射T在Σ_1上间断。图10显示了Λ是连通但非凸的情形。Ω是矩形、Λ是哑铃形、密度函数是常数、OT映射是不连续的、奇异点集合$\Sigma_1 = \gamma_1 \cup \gamma_2$。

图11显示了$\mathbb{R}^3$中两个概率测度之间的OT映射。源测度μ和目标测度υ均为均匀分布，Ω的支集是单位实心球，Λ的支集是实心斯坦福（Stanford）兔子。我们基于Theorem 4.2计算了Brenier势能$u:\Omega\to\mathrm{R}$。为了可视化映射，我们按如下方式插值概率测度：

$$\rho_t := [(1-t)\mathrm{id} + t\nabla u]_\#\mu,\ 0 \leqslant t \leqslant 1$$

图11显示了插值测度ρ_t的支集。表面的褶皱是奇异点集，其中OT映射是不连续的。

在一般情况下，由于实际数据分布、嵌入流形Σ以



(a) (b)

图9. 不连续的OT映射，由基于Theorem 4.2的一个GPU算法实现生成。（a）源域；（b）目标域。（a）图中间的线代表的是奇异点集合Σ_1。

及编码和解码映射的复杂性，目标测度支集很少是凸的，所以传输映射几乎不可能整体上都连续。

另外，一般的DNN，如ReLU DNN只能是逼近连续映射。ReLU DNN所表示的函数空间不包含所需的非连续传输映射。训练过程，即搜索过程，将出现以下三种情况：

（1）训练过程不稳定、不收敛。

（2）搜索过程会收敛到Λ的多个连通分支之一，映射会收敛到所期望的传输映射的一个连续分支。这意味着我们遇到了模式崩溃。

（3）训练过程能使传输映射成功覆盖所有模式，但同时也覆盖了Λ以外的区域。在实际应用中，这种情况将导致GAN产生不真实的样本。如图12所示。因此，从理论上讲，直接使用DNN来近似OT映射是不可能的。

5.3. AE-OT 模型

如图4所示，我们将GAN的两个主要任务分为流形学习和概率分布变换。第一个任务是通过AE来计算编码映射f_θ和解码映射g_ξ；第二个任务是利用变分方法来计算隐空间中的OT映射T。编码映射f_θ将实际数据分布υ前推为$(f_\theta)_\#\upsilon$。在隐空间中，T将均匀分布μ映射到$(f_\theta)_\#\upsilon$。

AE-OT模型有许多优势。寻找OT映射实际上是一个凸优化问题，这保证了解的存在性和唯一性。训练过程是稳定的，并采用了拟牛顿法进行超线性收敛。未知数的数量与训练样本的数量相等，避免了过度参数化。并行OT映射算法可以通过使用GPU来实现。OT映射的误差限可以通过Monte Carlo方法中的采样密度来控制。具有自适应性的等级分层算法进一步提高了计算效率。另外，AE-OT模型可以消除模式崩溃。

6. 实验结果

在这一部分，我们将展示实验结果。

6.1. 训练过程

AE-OT模型的训练主要包括两个步骤，即训练AE和寻找OT映射。正如第4节所述，使用GPU的算法实现来完成OT的步骤。在训练AE过程中，我们使用Adam算法[49]来优化神经网络的参数，其中学习率为0.003，$\beta_1 = 0.5$，$\beta_2 = 0.999$。当L^2损失停止下降时，这意味着神经网络找到了良好的编码映射，我们固定编码器部分并继续训练神经网络以获得解码映射。编码器固定前后

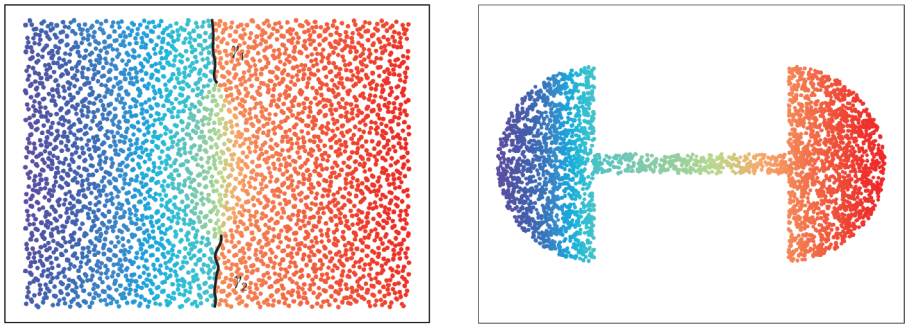


图10. 不连续的OT映射，由基于Theorem 4.2的一个GPU算法实现生成。（a）源域；（b）目标域。（a）图中的γ_1和γ_2是两个奇异点集合。

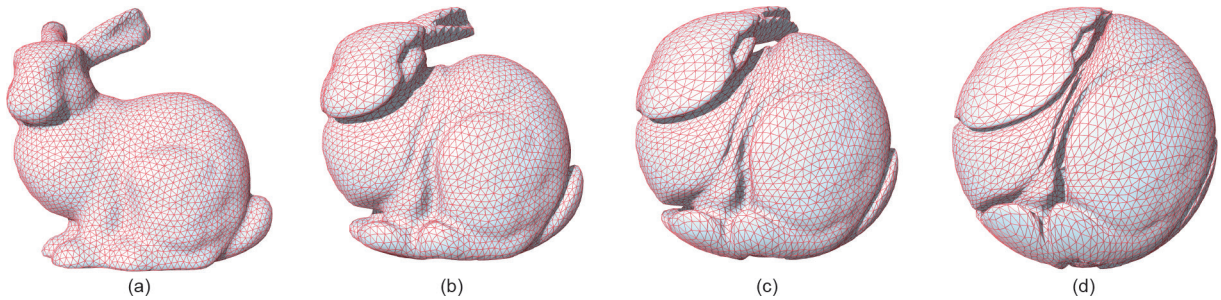


图11. 从Stanford兔子到实心球的OT映射。边界曲面上的皱褶是奇异点集合。（a）~（d）显示了变化过程。

(a) (b)

图12. AE-OT模型生成的人脸图像。（a）生成的实际人脸图像；（b）经过奇异点的路径。（b）图中心位置处的图像的传输映射是非连续的。

的训练损失见表1。接下来，为了找到从给定分布（我们在这里使用均匀分布）到隐空间特征的OT映射，我们从均匀分布中随机采样100N个随机样本点来计算能量梯度。这里，N是数据集隐空间特征的数目。实验中，θ_{ij}对于不同数据集也是不一样的。具体来说，对于MNIST和FASHION-MNIST两个数据集，θ_{ij}是0.75，但对于CIFAR10和CELEBA数据集，θ_{ij}分别为0.68和0.75。

我们的AE-OT模型是在Linux平台上通过使用PyTorch来实现的。所有实验均在GTX1080Ti上进行。

6.2. 传输映射不连续性测试

在这个实验中，我们的目的是去验证我们的假设，即在大多数实际应用中，目标测度的支集是非凸的、奇异点集是非空的以及在奇异点集上概率分布变换映射是非连续的。

如图12所示，我们使用AE来计算从CelebA数据集(Σ, υ)到隐空间Z的编码和解码映射，其中，编码映射$f_\theta: \Sigma \to Z$在隐空间上将υ前推为$(f_\theta)_\#\upsilon$。在隐空间中，我们用第4节所描述的算法计算了OT映射，即$T: Z \to Z$, 其中T将单位立方体ζ中的均匀分布映射为$(f_\theta)_\#\upsilon$。然后，我们从分布ζ中随机抽取样本z，并使用解码映射$g_\xi: Z \to \Sigma$将$T(z)$映射为生成的人脸图像$g_\xi \circ T(z)$。图12（a）展示了由该AE-OT模型生成的实际人脸图像。

如果隐空间中前推测度$(f_\theta)_\#\upsilon$的支集是非凸的，则存在奇异点集合Σ_k，其中$k>0$。我们希望验证Σ_K的存在。我们在隐空间的单位立方体中随机划上一条线段，然后沿着该线段密集插值以生成面部图像。如图12(b)所示，

表1 编码器固定前后AE的L^2损失

Situation	Dataset			
	MNIST	FANSION	CIFAR-10	CelebA
Before	0.0013	0.0026	0.0023	0.0077
After	0.0005	0.0011	0.0018	0.0074

我们找到了一条线段γ，并生成了一个变形序列，该序列的起点是具有一对棕色眼睛的男孩面部图像，终点是具有一对蓝色眼睛的女孩面部图像。在图像中间部分，我们生成的人脸的一只眼睛是蓝色的，另一只眼睛是棕色的。这些不真实人脸图像，应该在流形Σ之外。这意味着线段γ穿过了奇异点集Σ_k，而传输映射T在其上是不连续的，这也验证了我们的猜想是正确的，即被编码的人脸图像测度的支集在隐空间中是非凸的。

同时，我们发现AE-OT模型将训练速度提升了5倍，并且提高了模型的收敛稳定性，这是因为OT过程是一种凸优化过程。这为改进现有的GAN模型提供了一种很有前途的方法。

6.3. 模式崩溃比较

由于合成数据集由明确的分布和已知模式组成，因此利用这种数据集进行实验，可以精确地测量模式崩溃。我们选择了两个在之前的工作[50,51]中已经研究或提出的合成数据集——2D网格数据集。

关于模式崩溃测量指标的选择，我们选取了三种以前使用过的指标[50,51]。模式数量（number of modes）是指由生成模型生成的样本所捕捉到的模式个数。在这个指标中，如果在该模式的三个标准差范围内没有生成

样本，则我们判定该模式已失效。高质量样本的百分比（percentage of high-quality samples）测量的是在最近模式的三个标准差范围内生成的样本比例。参考文献[51]使用了第三个测量指标，即逆Kullback-Leibler（KL）散度。对于这个指标，每个生成样本都被分配给离其最近的模式，我们计算了被分配给每个模式的样本的直方图。该直方图形成了一种离散分布，然后我们计算了由真实数据形成的直方图的离散分布的KL散度。直观地说，该指标测量了生成样本在所有模式间关于真实分布的平衡程度。

在参考文献[51]中，作者用以上三种指标评估了GAN [26]、ALI [52]、MD [30]和PacGAN [51]在合成数据集上的表现。每个训练实验使用的生成器都具有相同的网络结构，训练参数共约400k个。网络训练的训练样本共有100k个，迭代次数为400次。对于AE-OT实验，由于源空间和目标空间都是2D，因此我们不需要训练任何AE。我们直接计算了单位正方形上的一致分布与真实数据分布之间的半离散OT映射。理论上，OT映射恢复所有模式所需的最小真实样本数量是每个模式需要一个真实样本。然而这可能导致在插值阶段生成低质量的样本。因此，对于OT映射的计算，我们采用了512个真实样本，并根据这个映射生成了新的样本。在这种情况下，我们注意到，在OT映射的计算中只有512个参数需要被优化，并且由于Hessian矩阵的正定性，优化过程是稳定的。我们的结果见表2，其中前面使用的方法的衡量标准见参考文献[51]。我们在合成数据集上的结果和GAN、PacGAN的结果如图13所示。

6.4. 与现有技术的比较

我们通过实验对本文提出的AE-OT模型和其他现有生成模型进行了比较，现有生成模型主要包括Lucic等在参考文献[33]中评估的对抗模型和Hoshen、Malik在参考文献[36]中研究的非对抗模型。

出于公平考虑，我们采用了相同的测试数据集和网络架构。数据集与参考文献[31,36]中的测试数据集类似，包括MNIST [53]、MNIST-Fansion [54]、CIFAR-10 [55]和CelebA [56]。网络架构与参考文献[33]中Lucic等使用的网络架构类似。特别是，在我们的AE-OT模型中，解码器的架构和参考文献[33]中GAN生成器的架构一样，并且编码器和解码器是对称的。

我们利用FID评分[31]和PRD曲线作为评估标准来比较我们的模型和现有的生成模型。FID评分衡量了生成结果的视觉保真度，并且对图象损坏具有鲁棒性。但是FID评分对模式的添加和删除非常敏感[33]，因此我们又使用了PRD曲线，PRD曲线可以量化真实数据集上模式丢失和添加的程度[32]。

6.4.1. 利用 FID 评分进行比较

FID评分计算方法如下：①通过运行inception网络[30]来提取生成图像和真实图像中有视觉意义的特征；②利用Gaussian分布来拟合真实图像和生成图像的分布；③用如下公式计算两个Gaussian分布之间的距离：

$$\mathrm{FID} = \| u_r - u_g \|_2^2 + T_r\left[\Sigma_r + \Sigma_g - 2\left(\Sigma_r\Sigma_g\right)^{1/2}\right] \quad (34)$$

表2 2D格点数据集上的模式崩溃比较

Method	Modes	Samples	Reverse KL
GAN	17.3 ± 0.8	94.8 ± 0.7%	0.70 ± 0.07
ALI	24.1 ± 0.4	95.7 ± 0.6%	0.14 ± 0.03
MD	23.8 ± 0.5	79.9 ± 3.2%	0.17 ± 0.03
PacGAN2	23.8 ± 0.7	91.3 ± 0.8%	0.13 ± 0.04
PacGAN3	24.6 ± 0.4	94.2 ± 0.4%	0.06 ± 0.02
PacGAN4	24.8 ± 0.2	93.6 ± 0.6%	0.04 ± 0.01
AE-OT	25.0 ± 0.0	99.8 ± 0.2%	0.007 ± 0.002

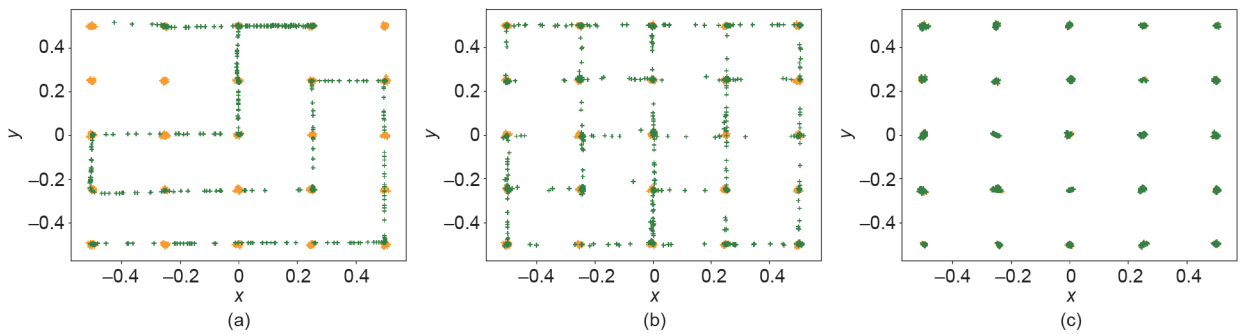


图13. 2D格点数据集上的模式崩溃比较。（a）GAN；（b）PacGAN4；（c）AE-OT。橙色点代表真实样本，绿色点代表生成样本。

式中，μ_r和μ_g分别代表真实分布的均值和生成分布的均值；Σ_r和Σ_g分别代表两个分布的方差。

比较的结果见表3和表4，几种GAN的统计数据来自Lucic等[33]，非对抗生成模型的统计数据则来自于Hoshen和Malik [36]。一般，我们提出的模型比其他现有生成模型能够获得更好的FID评分。

理论上来说，我们的AE-OT模型的FID评分和之前预训练的AE的FID评分接近，这从我们的实验中也得到了证实。

我们的AE采用的是Lucic等在参考文献[33]中提到的固定网络结构，它的性能不足以编码 CIFAR-10或者 CelebaA，因此我们必须下采样这些数据集。我们从CIFAR-10中随机选择了2.5×10^4张图像和从CelebaA中随机选择了1.0×10^4张图像来训练模型。即使是这样，我们的模型在CIFAR-10上依然取得了最好的FID评分。由于InfoGAN模型容量的有限性，CelebA的AE性能的FID评分（67.5）并不理想，这就使得生成的数据集的FID评分为68.4。通过在AE架构中增加两个额外的卷积层，CelebA的L^2损失将低于0.03，而且FID评分也超过了所有其他模型（28.6，如表4括号中所示）。

6.4.2. 利用 PRD 曲线进行比较

FID评分是度量生成分布和真实数据分布之间差距的一个有效方法，但它主要用于评价精确度，它不能准确地捕捉生成模型所能覆盖的真实数据比例。参考文献[32]中的方法将分布之间的散度分解为两个部分，即精确度和查全率。

给定一个参考分布P和一个学习分布Q，精确度可直观地衡量Q中样本的质量，而查全率衡量了Q所覆盖的P的比例。

我们使用Sajjadi等在参考文献[32]中介绍的$(F_8, F_{1/8})$的概念量化了精确度和查全率的相对重要程度。图14总结了对比结果。每个点代表的是一个有超参数集的具体模型。点离右上角越近，模型的性能越好。蓝色和绿色的点分别表示了参考文献[32]中评估的GAN和VAE，黄色的点代表的是参考文献[36]中的GLANN模型，而红色的点代表的是我们的AE-OT模型。

显然，在MNIST和FASHION-MNIST数据集上，我们提出的模型的性能要优于其他模型。对于CIFAR-10数据集，我们模型的精确度比GAN和GLANN的稍低，但是查全率是最高的。对于CelebA数据集，由于AE容量有限，我们的模型表现得不是很可观。但是，在AE里添加两个卷积层后，我们的模型得到了最高的评分。

6.4.3. 可视化比较

图15显示了由我们所提出的方法生成的图像和参考文献[33]中Lucic等研究的GAN以及参考文献[36]中Hoshen和Malik研究的非对抗模型生成的图像之间的可视化的比较结果。第一列是初始图像，第二列是由AE生成的结果，第三列是由Lucic等[33]采用GAN得到的

表3 用FID进行定量比较-I

Dataset	Adversarial				
	MM GAN	NS GAN	LSGAN	WGAN	BEGAN
MNIST	9.8	6.8	7.8	6.7	13.1
Fansion	29.6	26.5	30.7	21.5	22.9
CIFAR-10	72.7	58.5	87.1	55.2	71.4
CelebA	65.6	55.0	53.9	41.3	**38.9**

The best result is shown in bold. MM: manifold matching; NS: non-saturating; LSGAN: least squares GAN; BEGAN: boundary equilibrium GAN.

表4 用 FID 进行定量比较-II

Dataset	Non-Adversarial			Reference	
	VAE	GLO	GLANN	AE	AE-OT
MNIST	23.8	49.6	8.6	5.5	**6.4**
Fansion	58.7	57.7	13.0	4.7	**10.2**
CIFAR-10	155.7	65.4	46.5	28.2	**38.1**
CelebA	85.7	52.4	46.3	67.5	68.4 (**28.6**)

The best result is shown in bold.

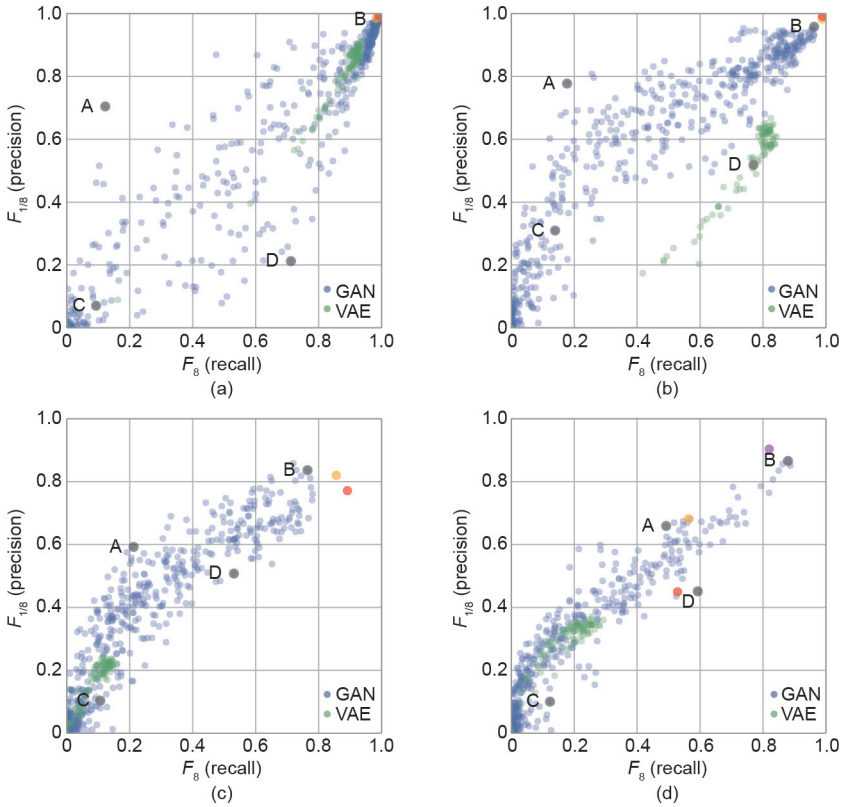


图14. 在四个数据集上，以$(F_8, F_{1/8})$的精确度-查全率进行比较。（a）MNIST；（b）FASHION；（c）CIFAR-10；（d）CelebA。黄褐色的点表示参考文献[36]中的结果。红色的点是利用本文所提出的方法生成的结果。（d）中紫色的点代表添加两个卷积层后，利用本文所提出的方法生成的结果。

图15. 生成图像质量在 4 个数据集上的可视化比较。第一列（a）是真实数据；第二列（b）是由AE生成的结果；第三列（c）显示的是由GAN[33]以最高的精确度-查全率$(F_8, F_{1/8})$生成的结果，它对应着图14中的B点；第四列（d）是参考文献[36]中的结果；最后一列（e）是利用本文所提出的方法生成的结果。

最好的生成结果，第四列是由Hoshen和Malik采用模型生成的结果，最后一列是用我们方法生成的结果。很明显，采用我们的方法生成了高质量的图像并且该图像包含了所有模式。

7. 结论

本文利用OT理论来解释GAN。根据数据流形分布假设，GAN主要完成两个任务——流形学习和概率分布变换。概率分布变换可以利用OT方法直接实现。OT理论解释了模式崩溃的基本原因，并指出生成器和判别器之间应该是合作而非竞争的内在关系。此外，我们提出了AE-OT模型，该模型提高了理论的严谨性、增强了训练的稳定性和效率，并且消除了模式崩溃问题。

我们的实验结果验证了我们的理论推测，即如果分布传输映射是不连续的，那么奇异点集的存在会导致模式崩溃。此外，通过将我们提出的模型与现有最先进的模型进行比较发现，我们提出的模型消除了模式崩溃，并在FID评分和PRD曲线方面要优于其他模型。

未来，我们将对流形学习阶段的理论理解进行探索，并用严格的方法使这部分黑匣子透明化。

致谢

本项目受到国家自然科学基金项目（61936002、61772105、61432003、61720106005和61772379）的资助。

Compliance with ethics guidelines

Na Lei, Dongsheng An, Yang Guo, Kehua Su, Shixia Liu, Zhongxuan Luo, Shing-Tung Yau, and Xianfeng Gu declare that they have no conflicts of interest or financial conflicts to disclose.

References

[1] Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks. In: Proceedings of the 34th International Conference on Machine Learning; 2017 Aug 6–11; Sydney, Australia; 2017 .p. 214–23.
[2] Tenenbaum JB, de Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science 2000;290(5500):2319–23.
[3] van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9(11):2579–605.
[4] Mescheder L, Geiger A, Nowozin S. Which training methods for GANs do actually converge? In: Proceedings of the 35th International Conference on Machine Learning; 2018 Jul 10–15; Stockholmsmässan, Sweden; 2018. p. 3478–87.
[5] Villani C. Optimal transport: old and new. Berlin: Springer Science & Business Media; 2008.
[6] Gu DX, Luo F, Sun J, Yau ST. Variational principles for Minkowski type problems, discrete optimal transport, and discrete Monge–Ampère equations. Asian J Math 2016;20(2):383–98.
[7] Peyré G, Cuturi M. Computational optimal transport. Found Trends Mach Learn 2019;11(5–6):355–607.
[8] Solomon J. Optimal transport on discrete domains. 2018. arXiv:1801.07745.
[9] Cuturi M. Sinkhorn distances: lightspeed computation of optimal transportation distances. Adv Neural Inf Process Syst 2013;26:2292–300.
[10] Solomon J, de Goes F, Peyré G, Cuturi M, Butscher A, Nguyen A, et al. Convolutional wasserstein distances: efficient optimal transportation on geometric domains. ACM Trans Graph 2015;34(4):66.
[11] Lei N, Su K, Cui L, Yau ST, Gu XD. A geometric view of optimal transportation and generative model. Comput Aided Geom Des 2019;68:1–21.
[12] Benamou JD, Brenier Y, Guittet K. The Monge-Kantorovitch mass transfer and its computational fluid mechanics formulation. Int J Numer Methods Fluids 2002;40(1–2):21–30.
[13] Jean-David Benamou BDF, Oberman AM. Numerical solution of the optimal transportation problem using the Monge-Ampère equation. J Comput Phys 2014;260:107–26.
[14] Nicolas P, Gabriel P, Oudet E. Optimal transport with proximal splitting. SIAM J Imaging Sci 2014;7(1):212–38.
[15] Bengio Y, Mesnil G, Dauphin Y, Rifai S. Better mixing via deep representations. In: Proceedings of the 30th International Conference on Machine Learning; 2013 Jun 16–21; Atlanta, GA, USA; 2013. p. 552–60.
[16] Salakhutdinov R, Larochelle H. Efficient learning of deep Boltzmann machines. In: Proceedings of the 13th International Conference on Artificial Intelligence and Statistics; 2010 May 13–15; Chia Laguna Resort, Italy; 2010. p. 693–700.
[17] Kingma DP, Welling M. Auto-encoding variational bayes. 2013. arXiv:1312.6114.
[18] Rezende DJ, Mohamed S, Wierstra D. Stochastic backpropagation and approximate inference in deep generative models. 2014. arXiv:1401.4082.
[19] Makhzani A, Shlens J, Jaitly N, Goodfellow I, Frey B. Adversarial autoencoders. 2015. arXiv:1511.05644.
[20] Tolstikhin I, Bousquet O, Gelly S, Schoelkopf B. Wasserstein auto-encoders. 2017. arXiv:1711.01558.
[21] He X, Yan S, Hu Y, Niyogi P, Zhang HJ. Face recognition using laplacianfaces. IEEE Trans Pattern Anal Mach Intell 2005;27(3):328–40.
[22] Arandjelovi'c O. Unfolding a face: from singular to manifold. In: Proceedings of the 9th Asian Conference on Computer Vision; 2009 Sep 23–27; Xi'an, China; 2009. p. 203–13.
[23] Salimans T, Karpathy A, Chen X, Kingma DP. PixelCNN++: Improving the PixelCNN with discretized logistic mixture likelihood and other modifications. 2017. arXiv:1701.05517.
[24] Oord Ad, Kalchbrenner N, Kavukcuoglu K. Pixel recurrent neural networks. 2016. arXiv:1601.06759.
[25] Van Den Oord A, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, et al. WaveNet: a generative model for raw audio. 2016. arXiv:1609.03499.
[26] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. 2014. arXiv:1406.2661.
[27] Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC. Improved training of wasserstein GANs. 2017. arXiv:1704.00028.
[28] Miyato T, Kataoka T, Koyama M, Yoshida Y. Spectral normalization for generative adversarial networks. 2018. arXiv:1802.05957.
[29] Zoran D, Weiss Y. From learning models of natural image patches to whole image restoration. In: Proceedings of the 2011 International Conference on Computer Vision; 2011 Jun 6–11; Barcelona, Spain; 2011. p. 479–86.
[30] Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X. Improved techniques for training GANs. 2016. arXiv:1606.03498.
[31] Heusel M, Ramsauer H, Unterthiner T, Nessler B, Klambauer G, Hochreiter S. GANs trained by a two time-scale update rule converge to a Nash equilibrium. 2017. arXiv:1706.08500.
[32] Sajjadi MS, Bachem O, Lucic M, Bousquet O, Gelly S. Assessing generative models via precision and recall. 2018. arXiv:1806.00035.
[33] Lucic M, Kurach K, Michalski M, Gelly S, Bousquet O. Are GANs created equal? A large-scale study. 2018. arXiv:1711.10337.
[34] Bojanowski P, Joulin A, Lopez-Paz D, Szlam A. Optimizing the latent space of generative networks. 2017. arXiv:1707.05776.
[35] Li K, Malik J. Implicit maximum likelihood estimation. 2018. arXiv:1809.09087.
[36] Hoshen Y, Malik J. Non-adversarial image synthesis with generative latent nearest neighbors. 2018. arXiv:1812.08985.
[37] Dinh L, Krueger D, Bengio Y. NICE: non-linear independent components estimation. 2014. arXiv:1410.8516.
[38] Dinh L, Sohl-Dickstein J, Bengio S. Density estimation using real NVP. 2017. arXiv:1605.08803.
[39] Kingma DP, Dhariwal P. Glow: generative flow with invertible 1 × 1 convolutions. 2018. arXiv:1807.03039.
[40] LeCun Y, Chopra S, Hadsell R, Ranzota MA, Huang FJ. A tutorial on energy-based learning. In: Bakir G, Hofman T, Schölkopf T, Smola A, Taskar B, editors.

Predicting structured data. Cambridge, MA: The MIT Press; 2006.
[41] Dai J, Lu Y, Wu Y. Generative modeling of convolutional neural networks. In: Proceedings of the 3rd International Conference on Learning Representations; 2015 May 7–9; San Diego, CA, USA; 2015.
[42] Nijkamp E, Hill M, Zhu S, Wu Y. On learning non-convergent non-persistent short-run MCMC toward energy-based model. 2019. arXiv:1904.09770.
[43] Bonnotte N. From Knothe's rearrangement to Brenier's optimal transport map. SIAM J Math Anal 2013;45(1):64–87.
[44] Brenier Y. Polar factorization and monotone rearrangement of vector-valued functions. Commun Pure Appl Math 1991;44(4):375–417.
[45] Caffarelli L. Some regularity properties of solutions of Monge Ampère equation. Commun Pure Appl Math 1991;44(8–9):965–9.
[46] Alexandrov AD. Convex polyhedra. New York: Springer New York; 2005.
[47] Guo X, Hong J, Lin T, Yang N. Relaxed wasserstein with applications to GANs. 2017. arXiv:1705.07164.
[48] Lei N, Guo Y, An D, Qi X, Luo Z, Gu X, et al. Mode collapse and regularity of optimal transportation maps. 2019. arXiv:1902.02934.
[49] Kingma DP, Ba J. Adam: a method for stochastic optimization. 2014. arXiv:1412.6980.
[50] Srivastava A, Valkov L, Russell C, Gutmann MU, Sutton C. VeeGAN: reducing mode collapse in GANs using implicit variational learning. 2017. arXiv:1705.17761.
[51] Lin Z, Khetan A, Fanti G, Oh S. PacGAN: the power of two samples in generative adversarial networks. 2017. arXiv:1712.04086.
[52] Dumoulin V, Belghazi I, Poole B, Mastropietro O, Lamb A, Arjovsky M, et al. Adversarially learned inference. 2016. arXiv:1606.00704.
[53] LeCun Y, Cortes C, Burges CJC. The MNIST database of handwritten digits Available from: http://yann.lecun.com/exdb/mnist/ (2010).
[54] Xiao H, Rasul F, Vollgraf R. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. 2017. arXiv:1708.07747.
[55] Krizhevsky A. Learning multiple layers of features from tiny images. Technical report. Toronto: University of Toronto; 2009.
[56] Zhang Z, Luo P, Loy CC, Tang X. From facial expression recognition to interpersonal relation prediction. Int J Comput Vis 2018;126(5):550–69.