南華大學
UNIVERSITY OF SOUTH CHINA

毕业论文/设计外文翻译

题　　目	对深度学习的几何理解
学院名称	数理学院
指导老师	高有
职　　称	讲师
班　　级	信计 1802
学　　号	20184390213
学生姓名	朱柳承

2022 年 1 月 15 日

对深度学习的几何理解

摘要：这篇论文 [1] 详细地给出了在深度学习中如何运用最优传输理论的相关知识的方法与技巧。我参照顾险峰教授的课程以及相关资料进行了简单的翻译。论文翻译如下。

这项工作引入了生成对抗网络 (GAN) 的最优运输 (OT) 映射的概念。自然数据集具有内在结构，可以概括为流形分布原理：一类数据的分布接近于一个低维流形。GAN 主要完成两个任务：流形学习和概率分布转换。后者可以用经典的 OT 方法来进行。从 OT 的角度来看，生成器计算 OT 映射，而判别器计算生成的数据分布和真实数据分布之间的 Wasserstein 距离；两者都可以简化为凸几何优化过程。此外，OT 理论还发现了发生器和判别器之间存在内在的协作关系，而不是竞争关系，以及模态崩溃的根本原因。我们还提出了一种新的生成模型，该模型使用自编码器 (AE) 进行流形学习，并使用 OT 映射进行概率分布变换。该 AE-OT 模型提高了理论的严谨性和透明度，以及计算的稳定性和效率；特别是，它消除了模式崩溃。实验结果验证了我们的假设，并证明了我们所提出的模型的优点。

关键词：生成；对抗；深度学习；最优传输；模式崩溃

1 引言

生成对抗网络 (GAN) 已成为无条件图像生成的主要方法之一。当在多个数据集上进行训练时，GAN 能够产生逼真的和视觉上吸引人的样本。GAN 方法训练了一个从随机噪声中回归真实图像的无条件的生成器，以及一个测量生成的样本和真实图像之间的差异的鉴别器。GAN 已经经过多次改进。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 数据库的流形结构。每个手写数字图像具有尺寸 28×28，并被视为 $\mathbb{R}^{784}$ 图像空间中的一个点。MNIST 数据库集中在一个低维（2 维流形）附近。通过使用 t-SNE 流形嵌入算法 [3]，MNIST 数据集被映射到一个平面域上，每个图像被映射到一个点上。代表相同数字的图像被映射到同一个簇上，10 个簇分别被不同的颜色编码。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。

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。

1.2 GAN 理论模型

图 2 为 GAN 的理论模型。真实的数据分布 ν 主要集中在被嵌入背景空间 chi 中的流形 $\sum$ 上。$(\sum, \nu)$ 一起显示了真实数据集的内在结构。GAN 模型计算一个从隐空间 Z 到流形 $\sum$ 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 ν_θ。判别器计算了真实数据分布 ν 和生成数据分布 μ_θ 之间的距离，如 Wasserstein 距离 $W_c(\mu_\theta, \nu)$，其等价于 Kontarovich 势能 φ_ξ。

图 2: GAN 的理论模型。G：生成器；D：判别器。

尽管 GAN 有许多优点，但它们也有严重的缺点。在理论上，我们对深度学习的基本原则的理解仍然是原始的。在实践中，GAN 的训练是棘手的，其对超参数是敏感的；而且 GAN 经常会遇到模式崩溃问题。最近，Meschede 等人 [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 为生成器映射 $g_{theta} = h \circ T$ 的分解，其中 $h : Z \to \sum$ 为从隐空间到背景空间中的数据流形 $\sum$ 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。

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。$T_{\#}\chi$ 是由 T 推导出的前推测度。

1.3 OT 观点

OT 理论 [5] 研究了以最经济的方式将一种概率分布转化为另一种概率分布的问题。OT 提供了严格而强大的方法来计算最优映射，这些方法将一个概率分布转换为另一个分布，同时计算出它们之间的距离 [6]。

如前所述，GAN 完成了两个主要任务：流形学习和概率分布变换。后一项任务可以直接通过 OT 方法完成。具体地说，在图 3 中，可以用 OT 理论计算出概率分布变换映射 T。判别器计算了生成数据分布与真实数据分布之间的 Wasserstein 距离 $W_c(\mu_\theta, \nu)$，可以用 OT 方法直接计算。

从理论的角度来看，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 自动编码器-最优传输模型

为了降低 GAN 的训练难度，特别是为了避免模式崩溃，我们提出了一个基于 OT 理论的更简单的生成模型：一个自动编码器 (AE)OT 模型 (AE-OT)，如图 4 所示。

图 4: 生成模型 AE-OT，将 AE 和 OT 相结合。

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 $f_\theta : Z \to \sum$，和解码映射 $g_\xi : \sum \to Z$，目的是用于流形学习。OT 映射 $T : Z \to Z$，将白噪声 ζ 转换为由编码映射 $(f_\theta)_\# \nu$ 前推的数据分布。

AE-OT 模型有很多优点。从理论的角度来看，OT 理论已经建立和得到了人们充分的理解。通过对解码映射和 OT 映射进行解耦，可以提高生成模型的理论严谨性，使一部分黑盒透明。在实际应用中，将 OT 映射简化为一个凸优化问题，保证了解的存在性和唯一性，同时使训练过程不会陷入局部最优状态。与 OT 映射相关的凸能量具有显式的 Hessian 矩阵结构；因此，我们可以采用二阶收敛的牛顿方法，或采用超线性收敛的拟牛顿方法进行优化。相比之下，目前的生成模型基于线性收敛的梯度下降法。而且在 AE-OT 模型中，未知数个数与训练样本个数相同，从而避免了过参数化问题。在 Monte Carlo 方法中，采样密度可以完全控制 OT 映射的误差范围。具有自适应的层次算法进一步提高了效率。并行 OT 映射算法可以使用图形处理单元 (GPU) 实现。更重要的是，AE-OT 模型可以

消除模式崩溃问题。

1.5 贡献

本工作利用 OT 理论解释了 GAN 模型。GAN 可以完成两个主要任务：流形学习和概率分布变换。后一种任务可以使用 OT 方法来完成。生成器计算了 OT 映射，而判别器计算生成的数据分布和真实数据分布之间的 Wasserstein 距离。利用 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 节简要回顾了 OT 和 GAN 中相关的工作；第 3 节简要介绍了 OT 的基本理论和 Monge-Ampère 方程的正则性理论；第 4 节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 节从 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 问题在各种领域中都起着重要的作用。有关详细的概述，读者可以参照参考文献 [7] 和 [8]。

当输入域和输出域都是 Dirac 质量时，OT 问题可以作为一个标准的线性规划 (LP) 任务来处理。为了将这个问题扩展到大的数据集，参考文献 [9] 的作者在原来的 LP 问题中添加了一个熵正则化器，则用 Sinkhorn 算法可以快速计算正则化问题。后来 Solomon 等人 [10] 随后通过引入快速卷积，提高了计算效率。

求解 OT 问题的第二种方法是通过 OT 问题和凸几何之间的联系来最小化凸能量 [6]，从而计算连续测度和逐点测度之间的 OT 映射。在参考文献 [11] 中，作者通过 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，则求解 OT 问题等价于求解著名的 Monge-Ampère 方程，这是一个高度非线性的椭圆型偏微分方程 (PDE)。有了一个额外的虚拟时间维数，这个问题可以通过计算流体动力学 [12–14] 来解决。

2.2 生成模型

在机器学习领域，能够生成复杂和高维数据的生成模型，最近正变得越来越重要。具体来说，生成模型主要用于从给定的图像数据集中生成新的图像。一些方法，包括深度信念网络 [15] 和深度玻尔兹曼机器 [16] 已经在早期阶段被引入。然而，这些方法的训练通常是棘手的和低效的。后来，从变分 AE(VAE)[17] 的方案中取得了一个巨大的突破，其中解码器使用变分方法 [17,18] 从 Gaussian 分布中逼近了真实的数据分布。最近遵循该方案的各种工作已经被提出，包括对偶自编码器（AAE）[19] 和 Wasserstein AE（WAE）[20]。虽然 VAE 的训练相对简单，但它们生成的图像看起来很模糊。在某种程度上，这是因为显式表示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，研究人员还逐步提出了其他非对抗性训练方法，包括 PixelCNN[23]、PixelRNN[24] 和 WaveNet[25]。然而，由于这些方法的自回归性质，新样本的生成无法并行。

2.3 对抗生成模型

针对上述模型的缺点，提出了 GAN[26]。尽管 GAN 是生成逼真样本的强大工具，但它们很难训练，并会出现模式崩溃。为了更好地进行 GAN 训练，人们提出了各种改进，包括改变损失函数 (如 WGAN[1]) 和通过剪切 [1]、梯度正则化 [4,27] 或光谱归一化 [28] 来正则化判别器。然而，GAN 的训练仍然很棘手，需

要仔细的超参数选择。

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 (特别是 GAN) 不适合接受这样的评估。传统上，GAN 的评估依赖于对少数例子或用户研究的视觉检查。近年来，人们提出了若干定量评价标准。Inception score(IS)[30] 同时衡量了多样性和图像质量。然而，它并不是一个距离度量标准。为了克服 IS 的缺点，研究人员在参考文献中引入了 Fréchet inception distance(FID)[31]。FID 已被证明对图像损坏具有鲁棒性，并与视觉保真度很好地相关性。在最近的研究 [32] 中，引入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。生成潜在优化 (GLO)[34] 采用了一种“无编码 AE”方法，其中生成模型用非对抗损失函数训练，获得比 VAE 更好的结果。隐式极大似然估计 (IMLE)[35] 提出了一种与迭代最近点 (ICP) 相关的生成模型训练方法。后来，Hoshen 和 Malik[36] 提出生成隐含最近邻 (GLANN)，该方法结合 GLO 和 GLANN 的优势，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。

其他方法利用具有可控 Jacobian 矩阵的 DNN 直接逼近从噪声空间到图像空间的分布变换映射 [37-39]。最近，人们选择了基于能量的模型 [40–42]，通过用 [40–42] 表示能量函数，通过 Gibb 分布来模拟图像的分布。这些方法利用现有的模型交替生成假样本，然后利用生成的假样本和真实样本对模型参数进行优化。

3 最优传输理论

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 d 维欧几里得空间 $\mathbb{R}^d$ 的两个子集，μ 和 ν 分别是定义在 X 和 Y 上的两个概率测度，密度函数如下：

$$\mu(x) = f(x)dx$$
$$\nu(y) = g(y)dy$$

假设总测度相等，$\mu(X) = \nu(Y)$；即

$$\int_X f(x)dx = \int_Y g(y)dy \tag{1}$$

我们只考虑保测度的映射。

定义 3.1 (保留测度映射). 如果映射 $T : X \to Y$ 对于任何可测集合 $B \subset Y$，集合 $T^{-1}(B)$ 是 μ- 可测量的，并且 $[T^{-1}(B)] = \nu(B)$，那么映射 $T : X \to Y$ 是保测度的，即

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

测度保持条件记为 $T_{\#\mu} = \nu$，其中 $T_{\#\mu}$ 是由 T 诱导的前推测度。

给定一个成本函数 $c(x, y) : X \times Y \to \mathbb{R}$，它表示将每个单位质量从源移动到目标的成本，映射 $T : X \to Y$ 的总传输成本被定义为

$$C_t = \int_X c[x, T(x)]d\mu(x) \tag{3}$$

Monge 的 OT 问题来自于找到使总运输成本最小化的保测度映射。

问题 3.1 (Monge【43】;MP). 给定传输成本函数 $c(x, y) : X \times Y \to \mathbb{R}_{\geq 0}$，找

到保测度映射 $X \to Y$，使总传输成本最小化：

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

定义 3.2 (OT 映射). *Monge 问题的解决方案被称为 OT 映射。一个 OT 映射的总运输成本被称为 μ 和 ν 之间的 Wasserstein 距离，记为 $W_c(\mu, \nu)$。*

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

3.2 Kontarovich 的方法

根据成本函数及其测度的性质，在 $(X\ \mu)$ 和 $(Y\ \nu)$ 之间的 OT 映射可能不存在。Kontarovich 将传输映射扩展为传输方案，并定义了联合概率测测度 $\rho(x, y) : X \times Y \to \mathbb{R}_{\geq 0}$，使 ρ 的边缘概率分别等于 μ 和 ν。令投影映射形式化为 $\pi_x(x\ y) = x, \pi_y(x\ y) = y$，然后定义联合测度类如下：

$$\Pi(x, y) = \left\{\rho(x, y) : X \times Y \to \mathbb{R} : (\pi_x)_{\#}\rho = \mu, (\pi_y)_{\#}\rho = \nu\right\} \tag{6}$$

问题 3.2 (Kontarovich;KP). *给定一个传输成本函数 $c(x, y) : X \times Y \to \mathbb{R}_{\geq 0}$，找到保测度映射 $X \to Y$，使总传输成本最小化：*

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

Kontarovich 问题 (KP) 可以用 LP 方法来求解。由于 LP 等式的对偶性，因此 Eq.(7)（KP 方程）可以重新表述为对偶性问题 (DP)，如下：

问题 3.3 (对偶;DP). *给定一个传输成本函数 $c(x, y) : X \times Y \to \mathbb{R}_{\geq 0}$，找到使总运输成本最小化的联合概率测度 $\rho(x, y) : X \times Y \to \mathbb{R}_{\geq 0}$*

$$(DP) \quad \max_{\varphi, \psi} \left[\int_X \varphi(x) d\mu + \int_Y \psi(y) d\nu : \varphi(x) + \psi(y) \geq c(x, y)\right] \tag{8}$$

等式 (8) 的最大值给出了 Wasserstein 距离。现有的 WGAN 模型大多是基于 L^1 代价函数下的对偶公式。

定义 3.3 (c-转换). $\varphi : X \to \mathbb{R}$ 的 *c*-变换被定义为 $\varphi^c : Y \to \mathbb{R}$:

$$\varphi^c(y) = \underset{x \in X}{inf} \left[c(x,y) - \varphi(x) \right] \tag{9}$$

然后，*DP* 可以重写如下：

$$(DP) \quad W_c(\mu,\nu) = \underset{\varphi}{max} \int_X \varphi(x) d\mu + \int_Y \varphi^c(y) d\nu \tag{10}$$

3.3 Brenier 方法

对于二次欧氏距离代价，用 Brenier[44] 证明了 OT 映射的存在性、唯一性和内在结构

定理 3.1 (Brenier【44】). 假设 X 和 Y 是欧几里得空间 $\mathbb{R}^d$ 的子集，运输代价是欧氏距离的平方 $c(x,y) = \frac{1}{2} \|x-y\|^2$。此外，$\mu$ 是绝对连续的，μ 和 ν 存在有限的二阶矩

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

然后存在一个凸函数 $\mu : X \to \mathbb{R}$，即所谓的 *Brenier* 势能，其梯度映射 $\bigtriangledown u$ 给出了 *Monge* 问题的解：

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

Brenier 势能是常熟范围内唯一不变的；因此，最佳的传输映射是唯一的。

假设 *Briener* 势是 C^2 光滑的，那么它就是以下 *Monge-Ampère* 方程的解：

$$det(\frac{\partial^2 u(x)}{\partial x_i \partial x_j}) = \frac{f(x)}{g \circ \bigtriangledown u(x)} \tag{13}$$

对于 $\mathbb{R}^d$ 中的 L^2 运输成本 $c(x,y) = \frac{1}{2} \|x-y\|^2$，*c*-变换和经典 *Legendre* 变换具有特殊的关系。

定义 3.4 (Legendre 变换). 给定一个函数 $\varphi : \mathbb{R}^n \to \mathbb{R}$，其 *Legendre* 变换定义如下：

$$\varphi^*(y) = \underset{x}{sup} \left[\langle x, y \rangle - \varphi(x) \right] \tag{14}$$

当 $c(x,y)=\frac{1}{2}\|x-y\|^2$ 可以证明，以下关系成立：

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

定理 3.2 (Briener 极分解【44】). 假设 X 和 Y 是欧几里得空间 $\mathbb{R}^d$，μ 是绝对连续的 *Lebesgue* 测度，和映射 $\varphi: X\to Y$ 前推 μ 为 ν，即 $\varphi_{\#}\mu=\nu$，然后存在一个凸函数 $u: X\to\mathbb{R}$，这样 $\varphi=\bigtriangledown u(x)\circ s$，在 $s: X\to X$ 是保测度的，$s_{\#}\mu=\mu$。此外，这个分解是唯一的。

以下定理在 OT 理论中是众所周知的

定理 3.3 (Villani【44】). 给定紧凸区域 $\Omega\subset\mathbb{R}^d$ 的 μ 和 ν，对于代价 $c(x,y)=h(x-y)$ 存在一个 *OT* 方案 ρ，且 h 严格凸。它是唯一的，并且形式为 $(id\ T_{\#})\mu$ *(id*：恒等映射*)*，前提是 μ 是绝对连续的，而 $\partial\Omega$ 可以忽略不计。此外，存在 *Kantorovich* 势能 φ，T 可以表示为：

$$T(x)=x-(\bigtriangledown h)^{-1}\left[\bigtriangledown\varphi(x)\right]$$

当 $c(x,y)=\frac{1}{2}\|x-y\|^2$，我们有

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

在这种情况下，*Brenier* 势在 μ 和 *Kantorovich* 势在 φ 有以下关系：

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

3.4 OT 映射的正则性

设 Ω 和 Λ 是 $\mathbb{R}^d$ 上的两个有界光滑开集，设 $\mu=fdx$ 和 $\nu=gdy$ 是该 $\mathbb{R}^d$ 上的两个概率测度 $f\mid_{\mathbb{R}^d\ \Omega}=0$ 和 $g\mid_{\mathbb{R}^d\ \Lambda}=0$。假设 f 和 g 在 Ω 和 Λ 上分别有界远离零和无穷大。

3.4.1 凸目标域

定义 3.5 (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 $|f(x) - f(y)| \leq C \|x - y\|^{\alpha}$ 对于 f 定义域中的所有 x 和 y 都成立。。

定义 3.6 (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 $k \geq 0$，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 $0 < \alpha \geq 1$。$C^{k,\alpha}_{loc}(\Omega)$ 意味着上述条件适用于 Ω 的任意紧子集。

定理 3.4 (Caffarelli【45】)**.** 如果 Λ 是凸的，那么 *Brenier* 势 u 是严格凸的；此外，

(1) 如果 $\lambda \geq f, g \geq \frac{1}{\lambda}$ 当 $\lambda > 0$，则 $u \in C^{1,\alpha}_{loc}(\Omega)$ 。

(2) 如果 $f \in C^{k,\alpha}_{loc}(\Omega)$ 和 $g \in C^{k,\alpha}_{loc}(\Lambda)$ 对于 $f, g > 0$，则 $u \in C^{k+2,\alpha}_{loc}(\Omega)$ 和 $(k \leq 0, \alpha \in (0,1))$ 。

3.4.2 非凸目标域

如果 Λ 不是凸的，并且存在光滑的 f 和 g 使得 $u \notin C^1(\Omega)$，那么 OT 映射 $\bigtriangledown u$ 在奇点上是不连续的。

定义 3.7 (次梯度)**.** 给定一个开集 $\Omega \subset \mathbb{R}^d$ 和一个凸函数 $u : \Omega \to \mathbb{R}$，对于 $x \in \Omega$，u 在 **x** 处的次梯度（次微分）定义如下：

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

很明显，$u(x)$ 是一个闭凸集。几何上，如果 $\rho \in u(x)$，那么超平面 $l_{x,p}(z) = u(x) + \langle p, z - x \rangle$ 在 x 从下面接触 u；也就是说，在 Ω 和 $l_{x,p}(x) = u(x)$ 上的 $l_{x,p} \leq u$，其中 $l_{x,p}$ 是 u 在 **x** 上的支撑平面。

如果 *Brenier* 势 u 的次梯度 $\partial u(x)$ 包含一个点，那么 u 在 **x** 上是可微的。我们根据点的子梯度的维数对点进行分类，并定义集合 $\sum_k(u) =$

$\{x \in \mathbb{R}^d \mid \dim[\partial u(x)] = k\}, k = 0, 1, 2, \cdots, d$。

很明显，$\sum_0(u)$ 是正则点的集合，而 $\sum_k(u)$，其中 $k > 0$，是奇异点的集合。我们还定义了在 **x** 处的可达的子梯度如下：

$$\partial_* u(x) = \left\{ \lim_{x \to \infty} \bigtriangledown u(x_k) \mid x_k \in \sum_0, x_k \to x \right\}$$

由此可知，次梯度等于可达次梯度的凸包，即：

$$\partial u(x) = Convex \quad hull \quad \bigtriangledown_* u(x)$$

定理 3.5 (正则性)**.** 设 Ω，Λ 是两个有界开集，设在 Ω 和 Λ 之外的两个为零的概率密度 $f, g : \mathbb{R}^d \to \mathbb{R}^+$，并且在 Ω 和 Λ 上分别从零和无穷远处间有界。用 $T = \bigtriangledown u : \Omega \to \Lambda$ 表示由定理 *(3.1)* 提供的 *OT* 映射。然后存在两个相对封闭的集 $\sum_\Omega \subset \Omega$ 和 $\sum_\Lambda \Lambda$ 与 $\sum_\Omega = \sum_\Lambda = 0$，这样 $T : \Omega \quad \sum_\Omega \to \Lambda \quad \sum_\Lambda$ 是某些 $\alpha > 0$ 的 $C^{0,\alpha}_{loc}$ 的同态类。

我们称 $\sum_\Omega$ 为 *OT* 映射的奇异集 $\bigtriangledown u : \Omega \to \Lambda$。图 *5* 说明了奇异集的结构，使用基于定理 *4.2* 的算法计算。我们得到的计算结果如下：

$$\sum_0 = \frac{\Omega}{\{\sum_1 \cup \sum_2\}}, \sum_1 = \bigcup_{k=0}^{3} \gamma_k, \sum_2 = \{x_0, x_1\}$$

x_0 的次梯度，$\partial u(x_0)$，是 Λ 的整个内孔，而 $\partial u(x_1)$ 是阴影三角形。对于 $\gamma_k(t)$ 上的每个点，$\partial u[\gamma_k(t)]$ 是 Λ 外的一条线段。x_1 是 γ_1, γ_2 和 γ_3 的分岔点。$\sum_1$ 和 $\sum_2$ 上的 *Brenier* 势是不可微的，它们上的 *OT* 映射 $\bigtriangledown u$ 是不连续的。

图 5: OT 映射的奇异点集结构。

4 计算算法

Brenier 定理可以直接推广到离散的情况下。在 GAN 模型中，源测度 μ 被定义为紧凸域 Ω 上的均匀（或高斯）分布；目标测度 ν 表示为经验测度，它是狄拉克测度的和：

$$\nu = \sum_{i=1}^{n} \nu_i \delta(y - y_i) \tag{17}$$

其中，$Y = \{y_1, y_2, \cdots, y_n\}$ 为训练样本，具有权重：$\sum_{i=1}^{n} \nu_i = \mu(\Omega)$；$\delta$ 为特征函数。

每个训练样本 y_i 对应于 Brenier 势的一个支撑平面，表示如下：

$$\pi_{h,i}(x) = \langle x, y_i \rangle + h_i \tag{18}$$

其中，支撑平面的截距（高度）h_i 是一个未知的变量。我们将所有的高度变量表示为 $h = (h_1, h_2, \cdots, h_n)$。

欧几里得空间中一族超平面的包络是一个超曲面，它在某一点上与族的每个成员相切，这些切点一起构成了整个包络。如图 6 所示，Brenier 势 $u_h : \Omega \to \mathbb{R}$ 是一个由 h 决定的分段线性凸函数，它是其所有支撑平面的上包络线：

$$u_h(x) = \max_{i=1}^{n} \left[\pi_{h,i}(x)\right] = \max_{i=1}^{n} \left[\langle x, y_i \rangle + h_i\right] \tag{19}$$

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

Brenier 势能图是一个凸多面体。每个支撑平面 $\pi_{h,i}$，i 对应于多面体的一个

面。多面体的投影诱导了 Ω 的胞腔分解，其中每个支撑平面 $\pi_i(x)$ 投射到一个胞胞 $W_i(h), p$ 上是 $\mathbb{R}^d$ 中的任意点：

$$\Omega = \bigcup_{i=1}^{n} w_i(h) \cap \Omega, W_i(h) = \{p \in \mathbb{R}^d \mid \bigtriangledown u_h(p) = y_i\} \tag{20}$$

胞腔分解是一个功率图。$W_i \cap \Omega$ 的 μ-测度记为 $w_i(h)$：

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

梯度映射：$\bigtriangledown u_h : \Omega \to Y$ 将每个单元格 $W_i(h)$ 映射到一个点 y_i：

$$\bigtriangledown u_h : W_i(h) \to y_i, i = 1, 2, \cdots, n \tag{22}$$

给定等式 (17) 中的目标度量 ν，在等式 (19) 中存在一个离散的布雷尼尔势，其每个面 $w_i(h)$ 的投影 μ 体积等于给定的目标度量 ν_i。Alexandrov[46] 在凸几何中证明了这一点。

定理 4.1 (Alexandrov【46】). 假设 Ω 是一个紧凑的凸多面体，在 $\mathbb{R}^d$ 中内部为非空，$n_1, \cdots, n_k \subset \mathbb{R}^{n+1}$ 是不同的 k 个单位向量，$(n+1)$ 坐标为负，$v_1, \cdots, v_k > 0$，因此 $\sum_{i=1}^{k} \nu_i = vol(\Omega)$ 。然后存在一个有精确 k 余维-*1* 面 $F_1, \cdots, F_k$ 的凸多面体 $P \subset \mathbb{R}^{n+1}$，其中 n_i 是 F_i 的法向量，并且在 Ω 和 F_i 的投影之间有体积 v_i。此外，这种 P 在垂直平移之前是唯一的。

Alexandrov 的解的存在性证明是基于代数拓扑，这不是构造性的。最近，*Gu* 等人 *[6]* 提供了一个基于变分方法的构造性证明。

定理 4.2 (参考文献【6】). 设 μ 是定义在 $\mathbb{R}^d$ 紧凸域 Ω 上的概率测度，设 $Y = \{y_1, y_2, \cdots, y_n\}$ 是 $\mathbb{R}^d$ 中的一个不同点。然后对于 $\sum_{i=1}^{n} v_i = \mu(\Omega)$ 的任何 $v_1, v_2, \cdots, v_n > 0$，对于所有 i，存在唯一的添加一个常数 $(c, c, \cdots, c)$ 使得 $w_i(h) = v_i$ 的 $h = (h_1, h_2, \cdots, h_n) \in \mathbb{R}^n$。向量 **h** 是以下凸能量的唯一最小值参数：

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

定义在一个开的凸集上

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

此外，$\bigtriangledown u_k$ 使二次代价最小化

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

在所有的传输映射中，$T_{\#\mu} = \nu$。

上述凸能量在等式 *(23)* 中的梯度值如下：

$$\bigtriangledown E(h) = [w_1(h) - v_1, w_2(h) - v_2, \cdots, w_n(h) - v_n]^T \tag{26}$$

能量的第 i 行和第 j 列元素如下：

$$\frac{\partial w_i}{\partial h_j} = -\frac{\mu(W_i \cap W_j \cap \Omega)}{\|y_i - y_j\|}, \frac{\partial w_i}{\partial h_i} = \sum_{j \neq i} \frac{\partial w_i}{\partial h_j} \tag{27}$$

如图 *6* 所示，*Hessian* 矩阵有一个明确的几何解释。图*6(a)* 显示了离散的布雷尼尔势 u_h，而图*6(b)* 显示了其使用定义 *(3.3)* 的勒让德变换。勒让德变换可以用几何方法构造：对于每个支撑平面 $\pi_{h,i}$，我们构造对偶点 $\pi^*_{h,i}$；对偶点的凸包 $\{\pi^*_{h,1}, \pi^*_{h,2}, \cdots, \pi^*_{h,n}\}$ 是勒让德变换的图 u^*_h。

u^*_h 的投影得到了一个的三角剖分 $Y = \{y_1, y_2, \cdots, y_n\}$，即加权的德劳内三角剖分。如图*7*所示，等式 *(20)* 中的能量图和加权的德劳内三角剖分是庞加莱对偶的：如果在幂图中，$W_i(h)$ 和 $W_j(h)$ 在一个 $(d-1)$-维胞腔上相交，那么在加权的德劳内三角剖分中，y_i 与 y_j 连接。在等式 *(27)* 中的黑森矩阵的元素是能量图中 $(d-1)$ 单元的 μ 体积与加权德劳内三角剖分中双边的长度之间的比值。

传统的幂次图可以与上述定理密切相关。

定义 4.1 (power 距离)**.** 给定一个功率权重为 ψ_i 的点 $y_i \in \mathbb{R}^d$，*power* 距离如下：

$$pow(x, y_i) = \|x - y_i\|^2 - \psi_i \tag{28}$$

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

定义 4.2 (power 图)**.** 给定加权点 $(y_1, \psi_1), \cdots, (y_k, \psi_k)$，幂图为单元分解的 $\mathbb{R}^d$：

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

其中，每个单元格都是一个凸多面体：

$$W_i(\psi) = \left\{x \in \mathbb{R}^d \mid pow(x, y_j) \leq pow(x, y_j)\right\} \tag{30}$$

加权德劳内三角剖分，记为 $T(\psi)$，是幂图的庞加莱对偶；如果 $W_i(\psi) \cap W_j(\psi) \neq \phi$，那么在加权德劳内三角剖分中有一条边连接 y_i 和 y_j。注意，$pow(x, y_i) \leq pow(x, y_j)$ 等价于

$$\langle x, y_i \rangle + \frac{1}{2}(\psi_i - \|y_i\|^2) \geq \langle x, y_j \rangle + \frac{1}{2}(\psi_j - \|y_j\|^2) \tag{31}$$

让 $h_i = \frac{1}{2}(\psi_i - \|y_i\|^2)$；然后我们将其定义重写如下：

$$W_i(\psi) = \left\{x \in \mathbb{R}^d \mid \langle x, y_i \rangle + h_i \geq \langle x, y_j \rangle + h_j, \forall j\right\} \tag{32}$$

在实践中，我们的目标是计算离散的布雷尼尔的潜在等式 (19) 通过优化凸能量等式 (23). 对于低维情况，我们可以通过计算牛顿方法直接计算梯度等式 (26) 和黑森矩阵等式 (27)。对于深度学习应用，直接计算黑森矩阵是不可行的；相反，我们可以使用梯度下降法或具有超线性收敛性的拟牛顿法。梯度的关键是估计 μ 体积 $w_i(h)$。这可以用蒙特卡罗方法来实现：我们从分布 中抽取 n 个随机样本，并计算在 $W_i(h)$ 范围内的样本数量，这是收敛于 μ 体积的比值。这种

方法是纯并行的，可以使用 GPU 来实现。此外，我们还可以使用分层的方法来进一步提高效率：首先，我们将目标样本分类为聚类，并计算到聚类质量中心的 OT 映射；其次，对于每个聚类，我们计算从相应的细胞到聚类内原始目标样本的 OT 映射。

图 8: Brenier 势能函数的奇异点集与 OT 映射的间断点集。

为了避免模态坍塌，我们需要在 Ω 中找到奇异点集。如图8所示，目标狄拉克测度有两个簇；源是单位平面圆盘上的均匀分布。布雷尼尔势函数的图是一个中间有一个脊的凸多面体。脊在圆盘上的投影是奇异集 $\sum_1(u)$，最优映射在 $\sum_1$ 上是不连续的。在一般情况下，如果两个单元格 $W_i(h)$ 和 $W_j(h)$ 相邻，则我们计算法线与相应支撑平面之间的夹角：

$$\theta_{ij} = \frac{\langle x_i, y_j \rangle}{\|y_i\| \|y_j\|}$$

如果 θ_{ij} 大于一个阈值，则公共面 $W_i(h) \cap W_j(h)$ 在不连续奇异集合中。

5 GAN 和最优传输

OT 理论为 GANgan 奠定了理论基础。许多最近的工作，如 WGAN[1]，WGAN-GP[27]，和 RW-GAN[47]，使用瓦瑟斯坦距离来测量生成的数据分布和

真实数据分布之间的偏差。

从 OT 的角度来看，生成器和鉴别器的最优解以一种封闭的形式联系起来；因此，生成器和鉴别器应该协作而不是竞争。更多细节请参见参考文献。[11]. 此外，Monge-Ampère 解的正则性理论可以解释 GAN[48] 中的模态坍缩。

5.1 竞争与合作

WGAN[1] 的 OT 视图如图 2 所示。根据流形分布假设，真实数据分布 ν 接近于嵌入在环境空间 χ 中的流形 $\sum$。生成器计算从潜在空间 Z 到环境空间的解码映射 g_θ，并将白噪声 ζ（即高斯分布）转换为生成的分布 μ_θ。该鉴别器通过计算坎托罗维奇的潜在 φ_ζ，计算 μ_θ 与真实数据分布 ν，$W_c(\mu_\theta, \nu)$ 之间的瓦瑟斯坦距离。g_θ 和 φ_ζ 都是由 DNN 实现的。

在训练过程中，生成器改善 g_θ 通过 $(g_\theta)_\# \zeta$ 来更好地近似 ν；鉴别器改进了坎托洛维奇势 φ_ζ 以改进对瓦瑟斯坦距离的估计。生成器和鉴别器相互竞争，而不共享中间的计算结果。在 L^1 成本函数下，WGAN 的替代训练过程可以表述为期望的最小-最大优化：

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

但如果我们将代价函数改变为 L^2 距离，那么根据定理 (3.2)，在最优时，布雷尼尔的势 μ 和康塔洛维奇的势 φ 通过等式 (16) 的封闭形式 $u(x) = \frac{1}{2}\|x\|^2 - \varphi(x)$ 联系起来, 生成器追求 OT 映射 $\bigtriangledown u$；鉴别器计算 φ。因此，一旦生成器达到最优值，无需任何训练就可以得到鉴别器的最优解，反之亦然。

更详细地说，假设在第 k 次迭代时，生成器映射为 g_θ^k。鉴别器计算康塔洛维奇势 φ_ζ，给出当前生成的数据分布 $\left(g_\theta^k\right)_\# \zeta$ 和真实数据分布 ν 之间的瓦瑟斯坦距离；$\bigtriangledown u$ 给出从 $\left(g_\theta^k\right)_\# \zeta$ 到 ν 的 OT 映射。因此，我们得到了以下结果：

$$v = (\bigtriangledown u)_\# \left[\left(g_\theta^k\right)_\# \zeta\right] = \left(\bigtriangledown u \circ \left(g_\theta^k\right)_\# \zeta\right)_\# \zeta = \left[(id - \bigtriangledown_{\varphi_\zeta}) \circ g_\theta^k\right]_\# \zeta$$

这意味着可以通过以下方式更新生成器映射：

$$g_\theta^{k+1} = (id - \bigtriangledown\varphi_\zeta) \circ g_\theta^k \tag{33}$$

这一结论表明，原则上可以跳过发电机的训练过程；在实践中，通过共享中间的计算结果，可以大大提高效率。因此，在设计加纳人的架构时，协作优于竞争。

5.2 模态崩溃与正则性

虽然 GAN 对许多应用程序都很强大，但它们有关键的缺点：首先，GAN 的训练很棘手，对超参数敏感，难以收敛；第二，GAN 易产生模式崩溃；第三，GAN 可能产生不真实的样本。不收敛、模态崩溃和生成不真实样本的问题都可以用 OT 映射的正则性定理 (3.5) 来解释。

根据布雷尼尔的极分解，定理 (3.2)，任何保测度映射都可以分解为两个映射，其中一个是 OT 映射，它是 Monge-Ampère 方程的解。根据正则性定理 (3.5)，如果目标测度 ν 的支持 Λ 有多个连通分量——即如果 ν 有多个模，或者 Λ 是非凸的，那么 OT 映射 $T : \Omega \to \Lambda$ 在奇异集 $\sum_\Omega$ 上是不连续的。

图9 显示了多簇的情况：Λ 有两个连接的组件，其中 OT 映射 T 沿着 $\sum_1$ 是不连续的。图10显示，即使是 Λ 也是连通的，尽管是非凸的。Ω 是一个矩形，Λ 是一个哑铃形，密度函数是常数，OT 映射是不连续的，奇异集 $\sum_1 = \gamma_1 \cup \gamma_2$。

图 9: 不连续的 OT 映射，由基于 Theorem4.2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 $\sum_1$。

图11为 $\mathbb{R}^3$ 中两个概率测度值之间的 OT 图。源测度 μ 和目标测度 ν 都是均匀分布；Ω 的支持是单位固体球，而 Λ 的支持是固体斯坦福兔子。基于定理 4.2，

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。

我们计算了布雷尼尔势 $u : \Omega \to \mathbb{R}$。为了可视化映射，我们插值概率测量如下：

$$\rho_t := [(1-t)id + t\bigtriangledown\mu]_{\#}\,\mu, 0 \le t \le 1$$

图11为插值测量值 ρ_t 的支持度。表面上的折叠是奇点集，其中 OT 映射是不连续的。

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。

在一般情况下，由于真实数据分布、嵌入流形 $\sum$ 和编码/解码映射的复杂性，目标测度量的支持度很少是凸的；因此，运输映射不能是全局连续的。

另一方面，一般的 dnn，如 ReLUdnn，只能近似于连续映射。由 ReLUdnn 表示的函数空间不包含所需的不连续传输映射。培训过程，或者，同样地，搜索过程将导致三种替代的情况：

(1) 训练过程不稳定，且不收敛。

(2) 搜索收敛于 Λ 的多个连通分量之一，并且映射收敛于所需的运输映射的一个连续分支。这意味着会遇到一个模式崩溃。

(3) 在培训过程中产生了一个交通地图，它成功地覆盖了所有的模式，但也覆盖了 Λ 以外的区域。在实际应用中，这将导致产生不现实的样本的现象，如图12的中间帧所示。

因此，在理论上，不可能直接使用 DNN 来近似 OT 映射。

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。

5.3 AE-OT 模型

如图4所示，我们分离了 GANs 的两个主要任务：流形学习和概率分布变换。第一个任务由 AE 执行，计算编码/解码映射 f_θ, g_ζ；第二个任务是使用显式变分方法来计算潜在空间中的 OT 映射 T。真实的数据分布 v 由编码映射 f_θ 向前推进，诱导 $(f_\theta)_\#\nu$。在潜在空间中，T 将均匀分布 μ 映射到 $(f_\theta)_\#\nu$。

AE-OT 模型有许多优点。在本质上，寻找 OT 映射是一个凸优化问题；保证了解的存在性和唯一性。采用拟牛顿方法，训练过程稳定且具有超线性收敛性。未知数的数量等于训练样本的数量，避免了过度参数化。并行 OT 映射算法可以使用 GPU 来实现。用蒙特卡罗方法可以通过采样密度来控制 OT 图的误差界。具有自适应能力的分层算法进一步提高了算法的效率。特别是，AE-OT 模型可以消除模式崩溃。

6 试验结果

在本节中，我们将报告我们的实验结果

6.1 训练过程

AE-OT 模型的训练主要包括两个步骤：训练 AE 和寻找 OT 图。OT 步骤是使用该算法的 GPU 实现来完成的，如第 4 节所述。在 AE 步骤中，在训练过程中，我们采用 Adam 算法 [49] 对中性网络的参数进行优化，学习速率为

0.003 $\beta_1 = 0.5$ $\beta_2 = 0.999$。当 L^2 损失停止下降时，这意味着网络已经找到了一个良好的编码图，我们冻结编码器部分，并继续训练网络进行解码图。编码器冻结前后的训练损失见表 (1)。接下来，为了从给定的分布（这里使用均匀分布）到潜在特征的分布中找到 OT 映射，我们从均匀分布中随机抽取 $100N$ 个随机点来计算能量的梯度。这里，N 是数据集的潜在特征的数量。此外，在实验中，θ_{ij} 被设置为对不同的数据集的不同。具体来说，对于 MNIST 和 Fashion-MNIST 数据集，θ_{ij} 被设置为 0.75，而对于 CIFAR-10 和 CelebA 数据集，它分别被设置为 0.68 和 0.75。

表 1: 编码器冻结前后 AEs 的 L^2 损失

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

我们的 AE-OT 模型是在一个 Linux 平台上使用 PyTorch 实现的。所有实验均在 GTX1080Ti 上进行。

6.2 传输映射的不连续性试验

在这个实验中，我们想要检验我们的假设：在大多数实际应用中，目标测度的支持是非凸的，奇异集是非空的，概率分布映射沿奇异集是不连续的。

如图12所示，我们使用 AE 计算从 CelebA 数据集 $(\sum, \nu)$ 到潜在空间 Z 的编码/解码映射；编码映射 $f_\theta : \sum \to Z$ 在潜在空间上将 ν 向前推到 $(f_\theta)_\# \mu$。在潜在空间中，我们基于第 4 节，$T : Z \to Z$ 中描述的算法计算 OT 映射，其中 T 将单位立方体 ζ 中的均匀分布映射到 $(f_\theta)_\# \nu$。然后，我们从分布 ζ 中随机抽取一个样本 z，并使用解码映射 $g_\xi : Z \to \sum$ 将 $T(z)$ 映射到生成的人类面部图像 $g_\xi \circ T(z)$。图12(a) 展示了由该 AE-OT 框架生成的真实面部图像。

如果潜在空间中的推进测度 $(f_\theta)_\# \nu$ 的支持是非凸的，则会有一个奇异集 $\sum_k$，其中 $k > 0$。我们想检测出 $\sum_k$ 的存在。我们在潜在空间中随机抽取单位立方体中的线段，然后沿着这条线段进行密集插值，生成面部图像。如图12(b) 所

示，我们找到了一个线段 γ，并在一个有一双棕色眼睛的男孩和一个有一对蓝眼睛的女孩之间生成了一个变形序列。在中间，我们生成了一张有一只蓝眼睛和一只棕色眼睛的脸，这绝对是不现实的，而且在 $\sum$ 之外。这个结果意味着线段 γ 经过一个奇异集 $\sum_k$，其中运输图 T 是不连续的。这也表明了我们的假设是正确的：编码的人脸图像测量在潜在空间上的支持是非凸的。

作为一个副产品，我们发现这个 AE-OT 框架提高了训练速度的 5 倍，并提高了收敛稳定性，因为 OT 步骤是一个凸优化。因此，它为改进现有的 GAN 提供了一种很有前途的方法。

6.3 模式崩溃比较

由于合成数据集由显式分布和已知模态组成，因此可以准确地测量模态坍塌。我们选择了两个在之前的工作中研究或提出的合成数据集：一个二维网格数据集。

为了选择模态崩溃的测量度量，我们采用了三个以前使用的度量 [50,51]。模式的数量计算由生成模型产生的样本所捕获的模式的数量。在这个度量中，如果在该模式的三个标准差内没有产生样本，则被认为是丢失的。高质量样本的百分比衡量在最近模式的三个标准差内产生的样本的比例。第三个度量标准，在参考文献中使用。[51]，是反向回-莱布勒 (KL) 散度。在这个度量中，每个生成的样本被分配到其最近的模式，并且我们计算分配在每个模式上的样本的直方图。然后，该直方图形成一个离散分布，然后计算其与真实数据形成的直方图的 KL 散度。直观地说，这衡量了生成的样本在真实分布的所有模式之间的平衡程度

在参考文献中。[51]，作者用上述三个指标评估了 GAN[26]、ALI[52]、MD[30] 和 PacGAN[51]。每个实验都在相同的生成器架构下进行训练，总共有大约 400k 的训练参数。这些网络在 100k 个样本上被训练了 400 个时代。对于 AE-OT 实验，由于源空间和目标空间都是二维的，因此不需要训练 AE。我们直接计算了一个半离散的 OT，它映射在单位平方上的均匀分布和经验的真实数据分布之间。理论上，OT 恢复所有模式所需的最小真实样本量是每个模式一个样本。然而，这

可能会导致在插值过程中产生低质量的样品。因此，对于 OT 计算，我们取 512 个真实样本，并基于此映射生成新的样本。我们注意到，在这种情况下，在 OT 计算中只有 512 个参数需要优化，并且由于凸正定黑森的存在，优化过程是稳定的。我们的结果在表 (2) 中所示，以前的方法的基准测试是从参考文献中复制出来的。[51] 为了说明这一点，我们在图13中绘制了与 GAN 和 PAcGAN 一起绘制的合成数据集上的结果。

表 2: 二维网格数据集的模式折叠比较

	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 ±0.003
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。

6.4 与现有技术水平的比较

我们设计了实验来比较我们提出的 AE-OT 模型与最先进的生成模型，包括由 Lucic 等人评估的对抗性模型。[33]，以及 Hoshen 和 Malik 研究的非对抗性模型。[36].

为了进行公平的比较，我们使用了相同的测试数据集和网络体系结构。这些数据集包括 MNIST[53]、MNIST-Dashion[54]、CIFAR-10[55] 和 CelebA[56]，与在 Refs 中测试的数据类似。[31,36]. 该网络架构与 Lucic 等人在参考文献中使用的类似。[33]. 特别是，在我们的 AE-OT 模型中，解码器的网络结构与 Ref 中 GANs 的生成器相同。[33]，编码器与解码器对称。

我们使用 FID 评分 [31] 和 PRD 曲线作为评价标准，将我们的模型与最先进

的生成模型进行了比较。FID 评分衡量生成结果的视觉保真度，并对图像损坏具有鲁棒性。然而，FID 评分对模式的添加和删除 [33] 很敏感。因此，我们也使用了 PRD 曲线，它可以量化真实数据集 [32] 上的模式下降程度。

6.4.1 与 FID 评分的比较

FID 得分的计算方法如下：① 通过运行初始网络 [30] 提取生成的图像和真实图像在视觉上有意义的特征；② 用高斯分布拟合真实和生成的特征分布；③ 使用以下公式计算两个高斯分布之间的距离：

$$FID = \|u_r - u_g\|_2^2 + T_r\left[\sum_r + \sum_g - 2\left(\sum_r \sum_g\right)^{\frac{1}{2}}\right] \tag{34}$$

其中，μ_r 和 μ_g 分别表示真实分布和生成分布的平均值；$\sum_r$ 和 $\sum_g$ 表示这些分布的方差

比较结果汇总见表 (3) 和表 (4)。各种 GANs 的统计数据来自 Lucic 等人的 [33]，而非对抗性生成模型的统计数据来自 Hoshen 和 Malik[36]。总的来说，我们提出的模型比其他最先进的生成模型获得了更好的 FID 分数。

表 3: 与 FID-I 进行的定量比较

Dataset	Adversarial				
	MM GAN	NS GAN	LS GAN	WGAN	BEGAN
MNIST	9.8	6.8	7.8	6.7	13.1
Fashion-MNIST	29.6	26.5	30.7	21.5	22.9
CIF AR-10	72.7	58.5	87.1	55.2	71.4
CelebA	65.6	55.0	53.8	41.3	**38.9**

最佳结果以粗体显示。

表 4: 与 FID-II 之间的定量比较

Dataset	Non-adversarial			Reference	
	VAE	GLO	CLANN	AE	AE-OT
MNIST	23.8	49.6	8.6	5.5	**6.4**
Fashion-MNIST	58.7	57.7	13.0	4.7	**10.2**
CIF AR-10	155.7	65.4	46.5	28.2	**38.1**
CelebA	85.7	52.4	46.3	67.5	68.4(**28.6**)

最佳结果以粗体显示。

理论上，我们的 AE-OT 模型的 FID 分数应该与预先训练过的 AEs 模型的分数很接近；我们的实验也验证了这一点。

我们的 AE 的固定网络架构采用了 Lucic 等人的 [33]；它的容量不足以编码 CIFAR-10 或 CelebaA，所以我们不得不对这些数据集进行降采样。我们从 CIFAR-10 中随机选择 25k 图像，从 CelebA 中选择 10k 图像来训练我们的模型。即便如此，我们的模型在 CIFAR-10 中获得了最好的 FID 分数。由于 InfoGAN 模型的有限容量，CelebA 的 AE 性能的 FID 为 67.5 并不理想，进一步导致生成的数据集的 FID 为 68.4。通过在 AE 体系结构中增加两个卷积层，CelebA 的 L^2 损失小于 0.03，FID 得分优于其他所有的模型（28.6，如表 (4) 的括号所示）。

图 14: 在四个数据集上，以 (F8, F1/8) 的精确度-召回率进行比较。（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。

6.4.2 与 PRD 曲线的比较

FID 评分是衡量生成分布与真实数据分布差异的有效方法，但它主要关注精度，不能准确地捕获生成模型可以覆盖的真实数据的哪个部分。参考文献中提出的方法。[32] 将分布之间的差异分解为两个组成部分：精度和召回率。

给定一个参考分布 P 和一个学习到的分布 Q，精度直观地衡量从 Q 中获得的样本的质量，而召回率则衡量 Q 所覆盖的 P 的比例。

我们使用了 Sajjadi 等人在参考文献中提出的 (F_8，$F_{1\,8}$) 的概念。[32] 来量化精度和查全率的相对重要性。图14总结了比较结果。每个点代表一个具有一组超参数的特定模型。一个点越靠近右上角，模型的性能就越好。蓝色和绿色的点表示参考文献中评估的 GANs 和 VAEs。[32]，黄点代表 Ref 中的 GLANN 模型。[36]，红点是我们的 AE-OT 模型.

很明显，我们提出的模型在 MNIST 和 Dashion-MNIST 上优于其他模型。对于 CIFAR-10 数据集，我们的模型的精度略低于 GANs 和 GLANN，但查全率最高。对于 CelebA 数据集，由于 AE 的容量有限，我们的模型的性能并不令人印象深刻。然而，在 AE 中增加了两个卷积层后，我们的模型获得了最好的分数。

6.4.3 视觉比较

图15显示了我们提出的方法生成的图像与 Lucic 等人研究的 GAN 生成的图像之间的视觉比较。[33] 和 Hoshen 和 Malik 研究的非对抗性模型。[36]. 第一列显示了原始图像，第二列显示了 AE 生成的结果，第三列显示最好的生成结果的差距在 Lucicetal.[33]，第四列显示所生成的结果霍申和马利克 [36] 的模型，和第五列显示我们的方法的结果。很明显，我们的方法可以生成高质量的图像，并覆盖了所有的模式。

7 结论

这项工作使用 OT 理论来解释 gan。根据数据流形分布假设，GANs 主要完成两个任务：流形学习和概率分布变换。后一种任务可以直接使用 OT 方法来完成。这一理论理解解释了模态崩溃的根本原因，并表明了发生器和鉴别器之间的内在关系应该是协作，而不是竞争。此外，我们提出了一个 AE-OT 模型，它提高了理论的严密性、训练的稳定性和效率，并消除了模式崩溃。

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。

我们的实验验证了我们的假设，即如果分布运输图是不连续的，那么奇点集的存在会导致模态崩溃。此外，当我们提出的模型与现有的技术进行比较时，我们的方法消除了模态崩溃，并在 FID 评分和 PRD 曲线方面优于其他模型。

在未来，我们将探索对流形学习阶段的理论理解，并使用一种严格的方法来使黑盒的这一部分透明。

致谢

本研究得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 conficts to disclose.

参考文献

[1] Na Lei, Dongsheng An, Yang Guo, Kehua Su, Shixia Liu, Zhongxuan Luo, Shing-Tung Yau, and Xianfeng Gu. A geometric understanding of deep learning. *Engineering*, 6(3):361–374, 2020.

注意：上述参考文献即本文的翻译原本。而文献中的参考文献，在翻译中没有给出，读者可以在附件中找到。

Journal Pre-proofs

Article

A Geometric Understanding of Deep Learning

Na Lei, Dongsheng An, Yang Guo, Kehua Su, Shixia Liu, Zhongxuan Luo, Shing-Tung Yau, Xianfeng Gu

PII: S2095-8099(19)30227-9
DOI: https://doi.org/10.1016/j.eng.2019.09.010
Reference: ENG 351

To appear in: *Engineering*

Received Date: 2 March 2019
Revised Date: 31 August 2019
Accepted Date: 11 September 2019

Please cite this article as: N. Lei, D. An, Y. Guo, K. Su, S. Liu, Z. Luo, S-T. Yau, X. Gu, A Geometric Understanding of Deep Learning, *Engineering* (2020), doi: https://doi.org/10.1016/j.eng.2019.09.010

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2020 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company.

Journal Pre-proofs

Research
Artificial Intelligence—Article

A Geometric Understanding of Deep Learning

Na Lei [a,#], Dongsheng An [b,#], Yang Guo [b], Kehua Su [c], Shixia Liu [d], Zhongxuan Luo [a], Shing-Tung Yau [e], Xianfeng Gu [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*

[#] These authors contributed equally to this work.

* Corresponding author.
E-mail address: gu@cs.stonybrook.edu (X. Gu).

ARTICLE INFO

Article history:
Received 2 March 2019
Revised 30 August 2019
Accepted 10 September 2019
Available online

Keywords:
Generative
Adversarial
Deep learning
Optimal transportation
Mode collapse

ABSTRACT

This work introduces an optimal transportation (OT) view of generative adversarial networks (GANs). Natural datasets have intrinsic patterns, which can be summarized as the manifold distribution principle: the distribution of a class of data is close to a low-dimensional manifold. GANs mainly accomplish two tasks: manifold learning and probability distribution transformation. The latter can be carried out using the classical OT method. From the OT perspective, the generator computes the OT map, while the discriminator computes the Wasserstein distance between the generated data distribution and the real data distribution; both can be reduced to a convex geometric optimization process. Furthermore, OT theory discovers the intrinsic collaborative—instead of competitive—relation between the generator and the discriminator, and the fundamental reason for mode collapse. We also propose a novel generative model, which uses an autoencoder (AE) for manifold learning and OT map for probability distribution transformation. This AE–OT model improves the theoretical rigor and transparency, as well as the computational stability and efficiency; in particular, it eliminates the mode collapse. The experimental results validate our hypothesis, and demonstrate the advantages of our proposed model.

1. Introduction

Generative adversarial networks (GANs) have emerged as one of the dominant approaches for unconditional image generation. When trained on several datasets, GANs are able to produce realistic and visually appealing samples. GAN methods train an unconditional generator that regresses real images from random noises and a discriminator that measures the difference between the generated samples and real images. GANs have received various improvements. One breakthrough was achieved by combing optimal transportation (OT) theory with GANs, such as the Wasserstein GAN (WGAN) [1]. In the WGAN framework, the generator computes the OT map from the white noise to the data distribution, and the discriminator computes the Wasserstein distance between the generated data distribution and the real data distribution.

1.1. Manifold distribution hypothesis

The great success of GANs can be explained by the fact that GANs effectively discover the intrinsic structures of real datasets, which can be formulated as the manifold distribution hypothesis: A specific class of natural data is concentrated on a low-dimensional manifold embedded in the high-dimensional background space [2].

Fig. 1 shows the manifold structure of the MNIST database. Each handwritten digit image has the dimensions 28 × 28, and is treated as a point in the image space $\mathbb{R}^{784}$. The MNIST database is concentrated close to a low-dimensional manifold. By using

2095-8099/© 2020 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Journal Pre-proofs

the t-SNE manifold embedding algorithm [3], the MNIST database is mapped onto a planar domain, and each image is mapped onto a single point. The images representing the same digit are mapped onto one cluster, and 10 clusters are color encoded. This demonstrates that the MNIST database is distributed close to a two-dimensional (2D) surface embedded in the unit cube in $\mathbb{R}^{784}$.

Fig. 1. Manifold distribution of the MNIST database. (a) Some handwritten digitals in MNIST database; (b) the embedded result of the digitals in two-dimensional (2D) plane by t-SNE algorithm. The x and y relative coordinates are normalized.

1.2. Theoretic model of GANs

Fig. 2 illustrates the theoretic model of GANs. The real data distribution v is concentrated on a manifold Σ embedded in the ambient space χ. (Σ, v) together show the intrinsic structure of the real datasets. A GAN model computes a generator map g_θ from the latent space Z to the manifold Σ, where θ represents the parameter of a deep neural network (DNN). ζ is a Gaussian distribution in the latent space, and g_θ pushes forward ζ to μ_θ. The discriminator calculates a distance between the real data distribution v and the generated distribution μ_θ, such as the Wasserstein distance $W_c(\mu_\theta, v)$, which is equivalent to the Kontarovich's potential φ_ξ.

Fig. 2: The theoretic model of GANs. G: generator; D: discriminator.

Despite GANs' advantages, they have critical drawbacks. In theory, the understanding of the fundamental principles of deep learning remains primitive. In practice, the training of GANs is tricky and sensitive to hyperparameters; GANs suffer from mode collapsing. Recently, Meschede et al. [4] studied nine different GAN models and variants showing that gradient-descent-based GAN optimization is not always locally convergent.

According to the manifold distribution hypothesis, a natural dataset can be represented as a probability distribution on a manifold. Therefore, GANs mainly accomplish two tasks: ① manifold learning—namely, computing the decoding/encoding maps between the latent space and the ambient space; and ② probability distribution transformation, either in the latent or image space, which involves transformation between the given white noise and the data distribution.

Fig. 3 shows the decomposition of the generator map $g_\theta = h \circ T$, where $h: Z \to \Sigma$ is the decoding map from the latent space to the data manifold Σ in the ambient space, the probability distribution transformation map T: $Z \to Z$. The decoding map h is for manifold learning, and the map T is for measure transportation.

Journal Pre-proofs

Fig. 3. The generator map is decomposed into a decoding map h and a transportation map T. $T_\#\zeta$ is the push-forward measure induced by T.

1.3. Optimal transportation view

OT theory [5] studies the problem of transforming one probability distribution into another distribution in the most economical way. OT provides rigorous and powerful ways to compute the optimal mapping to transform one probability distribution into another distribution, and to determine the distance between them [6].

As mentioned before, GANs accomplish two major tasks: manifold learning and probability distribution transformation. The latter task can be fully carried out by OT methods directly. In detail, in Fig. 3, the probability distribution transformation map T can be computed using OT theory. The discriminator computes the Wasserstein distance $W_c(\mu_\theta, v)$ between the generated data distribution and the real data distribution, which can be calculated directly using the OT method.

From the theoretical point of view, the OT interpretation of GANs makes part of the black box transparent, the probability distribution transformation is reduced to a convex optimization process using OT theory, the existence and uniqueness of the solution have theoretical guarantees, and the convergence rate and approximation accuracy are fully analyzed.

The OT interpretation also explains the fundamental reason for mode collapse. According to the regularity theory of the Monge–Ampère equation, the transportation map is discontinuous on some singular sets. However, DNN can only model continuous functions/mappings. Therefore, the target transportation mapping is outside of the functional space representable by GANs. This intrinsic conflict makes mode collapses unavoidable.

The OT interpretation also reveals a more complicated relation between the generator and the discriminator. In the current GAN models, the generator and the discriminator compete with each other without sharing the intermediate computational results. The OT theory shows that under the L^2 cost function, the optimal solution of the generator and the optimal solution of the discriminator can be expressed by each other in a closed form. Therefore, the competition between the generator and the discriminator should be replaced by collaboration, and the intermediate computation results should be shared to improve the efficiency.

1.4. Autoencoder–optimal transportation model

In order to reduce the training difficulty of GANs and, in particular, to avoid mode collapses, we propose a simpler generative model based on OT theory: an autoencoder (AE)–OT model, as shown Fig. 4.

As mentioned before, the two major tasks for generative models are manifold learning and probability distribution transformation. The AE computes the encoding map, $f_\theta : Z \to \Sigma$, and the decoding map g_ξ: $\Sigma \to Z$, for the purpose of manifold learning. The OT map, $T : Z \to Z$, transforms the white noise ζ to the data distribution push-forwarded by the encoding map, $(f_\theta)_\# v$.

Fig. 4. A generative model, the AE–OT model, combining an AE and an OT map.

Journal Pre-proofs

The AE–OT model has many merits. From a theoretical standpoint, the OT theory has been well established and is fully understood. By decoupling the decoding map and the OT map, it is possible to improve the theoretical rigor of generative models and make part of the black box transparent. In practice, the OT map is reduced to a convex optimization problem, the existence and the uniqueness of the solution are guaranteed, and the training process will not be trapped in the local optimum. The convex energy associated with the OT map has an explicit Hessian matrix; hence, the optimization can be performed using Newton's method with second-order convergence, or using quasi-Newton's method with superlinear convergence. In contrast, current generative models are based on the gradient descent method with linear convergence; the number of unknowns is equal to that of the training samples, in order to avoid the over-paramerization problem; the error bound of the OT map can be fully controlled by the sampling density in the Monte Carlo method; the hierarchical algorithm with self-adaptivity further improves the efficiency; and the parallel OT map algorithm can be implemented using a graphics processing unit (GPU). Most importantly, the AE–OT model can eliminate mode collapse.

1.5. Contributions

This work interprets the GAN model using OT theory. GANs can accomplish two major tasks: manifold learning and probability distribution transformation. The latter task can be carried out using OT methods. The generator computes an OT map, and the discriminator calculates the Wasserstein distance between the generated data distribution and the real data distribution. Using Brenier's theory, the competition between the generator and the discriminator can be replaced by collaboration; according to the regularity theory of the Monge–Ampère equation, the discontinuity of the transportation map causes mode collapse. We further propose to decouple manifold learning and probability distribution transformation by means of an AE–OT model, which makes part of the black box transparent, improves the training efficiency, and prevents mode collapse. The experimental results demonstrate the efficiency and efficacy of our method.

This paper is organized as follows: Section 2 briefly reviews the most related works in OT and GANs; Section 3 briefly introduces the fundamental theories in OT and the regularity theory of the Monge–Ampère equation; Section 4 introduces a variational framework for computing OT, which is suitable for a deep learning setting; Section 5 analyzes the GAN model from the OT perspective, explains the collaborative (instead of competitive) relation between the generator and the discriminator, and reveals the intrinsic reason for mode collapse; Section 6 reports the experimental results; and the paper concludes in Section 7.

2. Previous works

2.1. Optimal transportation

The OT problem plays an important role in various kinds of fields. For detailed overviews, we refer readers to Refs. [7] and [8].

When both the input and output domains are Dirac masses, the OT problem can be treated as a standard linear programming (LP) task. In order to extend the problem to a large dataset, the authors of Ref. [9] added an entropic regularizer into the original LP problem; as a result, the regularized problem could be quickly computed with the Sinkhorn algorithm. Solomon et al. [10] then improved the computational efficiency by the introduction of fast convolution.

The second type of method for solving the OT problem computes the OT map between continuous and point-wise measures by minimizing a convex energy [6] through the connection between the OT problem and convex geometry. In Ref. [11], the authors then linked the convex-geometry-viewed OT to the Kantorovich duality by means of the Legendre dual theory. The proposed method is an extension of this method in high dimension. If both the input and output are continuous densities, solving the OT problem is equivalent to solving the famous Monge–Ampère equation, which is a highly nonlinear elliptic partial differential equation (PDE). With an additional virtual time dimension, this problem can be relaxed through computational fluid dynamics [12−14].

2.2. Generative models

In the machine learning field, generative models, which are capable of generating complex and high-dimensional data, are recently becoming increasingly important and popular. To be specific, generative models are largely utilized to generate new images from given image datasets. Several methods, including deep belief networks [15] and deep Boltzmann machines [16] have been introduced in early stages. However, the training in these methods is generally tricky and inefficient. Later, a huge breakthrough was achieved from the scheme of variational AEs (VAEs) [17], where the decoders approximate real data distributions from a Gaussian distribution using a variational approach [17,18]. Various recent works following this scheme have been proposed, including adversarial AEs (AAEs) [19] and Wasserstein AEs (WAEs) [20]. Although VAEs are relatively simple to train, the images they generate look blurry. To some extent, this is because the explicitly expressed density functions may fail to represent the complexity of a real data distribution and learn the high-dimensional data distribution [21,22]. Other non-

Journal Pre-proofs

adversarial training methods have been proposed, including PixelCNN [23], PixelRNN [24], and WaveNet [25]. However, due to their auto-regressive nature, the generation of new samples cannot be paralleled.

2.3. Adversarial generative models

GANs [26] were proposed to solve the disadvantages of the above models. Although they are a powerful tool for generating realistic-looking samples, GANs can be difficult to train and suffer from mode collapsing. Various improvements have been proposed for better GAN training, including changing the loss function (e.g., the WGAN [1]) and regularizing the discriminators to be Lipschitz by clipping [1], gradient regularization [4,27], or spectral normalization [28]. However, the training of GANs is still tricky and requires careful hyperparameter selection.

2.4. Evaluation of generative models

The evaluation of generative models remains challenging. Early works include probabilistic criteria [29]. However, recent generative models (particularly GANs) are not amenable to such evaluation. Traditionally, the evaluation of GANs relies on visual inspection of a handful of examples or a user study. Recently, several quantitative evaluation criteria were proposed. The inception score (IS) [30] measures both diversity and image quality. However, it is not a distance metric. To overcome the shortcomings of the IS, the Fréchet inception distance (FID) was introduced in Ref. [31]. The FID has been shown to be robust to image corruption, and correlates well with visual fidelity. In a more recent work [32], precision and recall for distributions (PRD) was introduced to measure both precision and recall between generated data distribution and real data distribution. In order to fairly compare the GANs, a large-scale comparison was performed in Ref. [33], where seven different GANs and VAEs were compared under a uniform network architecture, and a common baseline for evaluation was established.

2.5. Non-adversarial methods

Various non-adversarial methods have also been proposed recently. Generative latent optimization (GLO) [34] employs an "encoder-less AE" approach in which a generative model is trained with a non-adversarial loss function, achieving better results than VAEs. Implicit maximum likelihood estimation (IMLE) [35] proposed an iterative closest points (ICP)-related generative model training approach. Later, Hoshen and Malik [36] proposed generative latent nearest neighbors (GLANN), which combines the advantages of GLO and GLANN, in which an embedding from the image space to latent space was first found using GLO, and then a transformation between an arbitrary distribution and latent code was computed using IMLE.

Other methods directly approximate the distribution transformation map from the noise space to the image space by means of DNNs with a controllable Jacobian matrix [37−39]. Recently, the energy-based models [40−42] have been chosen to model the image distribution through the Gibbs distribution by representing the energy function with DNNs. These methods alternatively generate fake samples using the current models, and then optimize the model parameters with the generated fake samples and real samples.

3. Optimal transportation theory

In this section, we introduce basic concepts and theorems in classic OT theory, with a focus on Brenier's approach and their generalization to the discrete setting. Details can be found in Villani's book [5].

3.1. Monge's problem

Suppose $X \subset \mathbb{R}^d, Y \subset \mathbb{R}^d$ are two subsets of d-dimensional Euclidean space $\mathbb{R}^d$, and μ and ν are two probability measures defined on X and Y, respectively, with the following density functions:

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

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

Suppose the total measures are equal, $\mu(X) = \nu(Y)$; that is

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

We only consider maps that preserve the measures.

Definition 3.1 (measure-preserving map). A map $T: X \to Y$ is measure preserving if for any measurable set $B \subset Y$, the set $T^{-1}(B)$ is μ-measurable and $\mu[T^{-1}(B)] = \nu(B)$, that is,

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

The measure-preserving condition is denoted as $T_{\#}\mu = \nu$, where $T_{\#}\mu$ is the push-forward measure induced by T.

Journal Pre-proofs

Given a cost function $c(\boldsymbol{x},\boldsymbol{y}): X\times Y\to\mathbb{R}_{\geq 0}$, which indicates the cost of moving each unit mass from the source to the target, the total transport cost of the map T: $X\to Y$ is defined to be

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

Monge's problem of OT arises from finding the measure-preserving map that minimizes the total transport cost.

Problem 3.2 (Monge's [43]; MP). Given a transport cost function $c(\boldsymbol{x},\boldsymbol{y}): X\times Y\to\mathbb{R}_{\geq 0}$, find the measure-preserving map T: $X\to Y$ that minimizes the total transport cost:

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

Definition 3.3 (OT map). The solutions to Monge's problem are called the OT maps. The total transportation cost of an OT map is called the Wasserstein distance between μ and ν, denoted as $W_c(\mu,\nu)$.

$$W_C(\mu,\nu)=\min_{T_{\#}\mu=\nu}\int_X c[\boldsymbol{x},T(\boldsymbol{x})]\mathrm{d}\mu(\boldsymbol{x}) \quad (5)$$

3.2. Kontarovich's approach

Depending on the cost function and measures, an OT map between (X,μ) and (Y,ν) may not exist. Kontarovich relaxed the transportation maps to transportation plans, and defined the joint probability measure $\rho(\boldsymbol{x},\boldsymbol{y})$: $X\times Y\to\mathbb{R}_{\geq 0}$, such that the marginal probability of ρ is equal to μ and ν, respectively. Let the projection maps formally be $\pi_{\boldsymbol{x}}(\boldsymbol{x},\boldsymbol{y})=\boldsymbol{x}$, $\pi_{\boldsymbol{y}}(\boldsymbol{x},\boldsymbol{y})=\boldsymbol{y}$, then define the joint measure class as follows:

$$\Pi(\mu,\nu)=\left\{\rho(\boldsymbol{x},\boldsymbol{y}): X\times Y\to\mathbb{R}:(\pi_{\boldsymbol{x}})_{\#}\rho=\mu,(\pi_{\boldsymbol{y}})_{\#}\rho=\nu\right\} \quad (6)$$

Problem 3.4 (Kontarovich; KP). Given a transport cost function $c(\boldsymbol{x},\boldsymbol{y}): X\times Y\to\mathbb{R}_{\geq 0}$, find the joint probability measure $\rho(\boldsymbol{x},\boldsymbol{y})$: $X\times Y\to\mathbb{R}_{\geq 0}$ that minimizes the total transport cost.

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

Kontarovich's problem (KP) can be solved using the LP method. Due to the duality of LP, Eq. (7) (the KP equation) can be reformulated as the duality problem (DP) as follows:

Problem 3.5 (duality; DP). Given a transport cost function $c(\boldsymbol{x},\boldsymbol{y}): X\times Y\to\mathbb{R}_{\geq 0}$, find the real functions φ: $X\to\mathbb{R}$ and ψ: $Y\to\mathbb{R}$, such that

$$(\mathrm{DP})\quad \max_{\varphi,\psi}\left[\int_X\varphi(\boldsymbol{x})\mathrm{d}\mu+\int_Y\psi(\boldsymbol{y})\mathrm{d}\nu:\varphi(\boldsymbol{x})+\psi(\boldsymbol{y})\leq c(\boldsymbol{x},\boldsymbol{y})\right] \quad (8)$$

The maximum value of Eq. (8) gives the Wasserstein distance. Most existing WGAN models are based on the duality formulation under the L^1 cost function.

Definition 3.6 (*c*-transformation). The c-transformation of φ: $X\to\mathbb{R}$ is defined as φ^c: $Y\to\mathbb{R}$:

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

The DP can then be rewritten as follows:

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

3.3. Brenier's approach

For the quadratic Euclidean distance cost, the existence, uniqueness, and intrinsic structure of the OT map were proven by Brenier [44].

Journal Pre-proofs

Journal Pre-proofs

Theorem 3.7 (Brenier's [44]). Suppose X and Y are subsets of the Euclidean space $\mathbb{R}^d$ and the transportation cost is the quadratic Euclidean distance $c(\boldsymbol{x},\boldsymbol{y}) = 1/2\|\boldsymbol{x}-\boldsymbol{y}\|^2$. Furthermore, μ is absolutely continuous and μ and ν have finite second-order moments

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

then there exists a convex function u: $X \to \mathbb{R}$, the so-called Brenier's potential, whose gradient map ∇u gives the solution to Monge's problem:

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

The Brenier's potential is unique up to a constant; hence, the optimal mass transportation map is unique.

Assuming that the Briener potential is C^2 smooth, then it is the solution to the following Monge–Ampère equation:

$$\det\left(\frac{\partial^2 u(\boldsymbol{x})}{\partial \boldsymbol{x}_i \partial \boldsymbol{x}_j}\right) = \frac{f(\boldsymbol{x})}{g \circ \nabla u(\boldsymbol{x})} \quad (13)$$

For the L^2 transportation cost $c(\boldsymbol{x},\boldsymbol{y}) = 1/2\|\boldsymbol{x}-\boldsymbol{y}\|^2$ in $\mathbb{R}^d$, the c-transform and the classical Legendre transform have special relations.

Definition 3.8 (Legendre transform). Given a function $\varphi : \mathbb{R}^n \to \mathbb{R}$, its Legendre transform is defined as follows:

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

It can be shown that the following relation holds when $c(\boldsymbol{x},\boldsymbol{y}) = 1/2\|\boldsymbol{x}-\boldsymbol{y}\|^2$:

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

Theorem 3.9 (Brenier's polar factorization [44]). Suppose X and Y are the Euclidean space $\mathbb{R}^d$, μ is absolutely continuous with respect to the Lebesgue measure, and a mapping φ: $X \to Y$ pushes μ forward to υ, $\varphi_{\#}\mu = \nu$, then there exists a convex function u: $X \to \mathbb{R}$, such that $\varphi = \nabla u \circ s$, where s: $X \to X$ is measure preserving, $s_{\#}\mu = \mu$. Furthermore, this factorization is unique.

The following theorem is well known in OT theory:

Theorem 3.10 (Villani [5]). Given μ and ν on a compact convex domain $\Omega \subset \mathbb{R}^d$, there exists an OT plan ρ for the cost $c(\boldsymbol{x}, \boldsymbol{y}) = h(\boldsymbol{x} - \boldsymbol{y})$, with h strictly convex. It is unique and of the form (id, $T_{\#}$)μ (id: identity map), provided that μ is absolutely continuous and $\partial\Omega$ is negligible. Moreover, there exists a Kantorovich's potential φ, and T can be represented as follows:

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

When $c(\boldsymbol{x},\boldsymbol{y}) = 1/2\|\boldsymbol{x}-\boldsymbol{y}\|^2$, we have

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

In this case, the Brenier's potential u and the Kantorovich's potential φ are related by the following:

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

3.4. Regularity of OT maps

Let Ω and Λ be two bounded smooth open sets in $\mathbb{R}^d$, and let $\mu = f\mathrm{d}\boldsymbol{x}$ and $\nu = g\mathrm{d}\boldsymbol{y}$ be two probability measures on $\mathbb{R}^d$ such that $f\big|_{\mathbb{R}^d\setminus\Omega} = 0$ and $g\big|_{\mathbb{R}^d\setminus\Lambda} = 0$. Assume that f and g are bounded away from zero and infinity on Ω and Λ, respectively.

3.4.1. Convex target domain

Journal Pre-proofs

Definition 3.11 (Hölder continuous). A real or complex-valued function f on a d-dimensional Euclidean space satisfies a Hölder condition, or is Hölder continuous, when there are nonnegative real constants C, $\alpha > 0$, such that $|f(\boldsymbol{x}) - f(\boldsymbol{y})| \le C\|\boldsymbol{x} - \boldsymbol{y}\|^{\alpha}$ for all x and y in the domain of f.

Definition 3.12 (Hölder space). The Hölder space $C^{k,\alpha}(\Omega)$, where Ω is an open subset of some Euclidean space and $k \ge 0$ is an integer, consists of those functions on Ω having continuous derivatives up to order k and such that the kth partial derivatives are Hölder continuous with exponent α, where $0 < \alpha \le 1$. $C^{k,\alpha}_{\text{loc}}(\Omega)$ means the above conditions hold on any compact subset of Ω.

Theorem 3.13 (Caffarelli [45]). If Λ is convex, then the Brenier's potential u is strictly convex; furthermore,

(1) If $\lambda \le f$, $g \le 1/\lambda$ for some $\lambda > 0$, then $u \in C^{1,\alpha}_{\text{loc}}(\Omega)$.

(2) If $f \in C^{k,\alpha}_{\text{loc}}(\Omega)$ and $g \in C^{k,\alpha}_{\text{loc}}(\Lambda)$, with $f, g > 0$, then $u \in C^{k+2,\alpha}_{\text{loc}}(\Omega)$ and $(k \ge 0, \alpha \in (0.1))$.

3.4.2. Non-convex target domain

If Λ is not convex and there exist f and g that are smooth such that $u \notin C^1(\Omega)$, then the OT map ∇u is discontinuous at singularities.

Definition 3.14 (subgradient). Given an open set $\Omega \subset \mathbb{R}^d$ and a convex function u: $\Omega \to \mathbb{R}$, for $\boldsymbol{x} \in \Omega$, the subgradient (subdifferential) of u at $\boldsymbol{x}$ is defined as follows:

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

It is obvious that $u(\boldsymbol{x})$ is a closed convex set. Geometrically, if $\boldsymbol{p} \in u(\boldsymbol{x})$, then the hyperplane $l_{\boldsymbol{x},\boldsymbol{p}}(\boldsymbol{z}) = u(\boldsymbol{x}) + \langle \boldsymbol{p}, \boldsymbol{z} - \boldsymbol{x} \rangle$ touches u from below at $\boldsymbol{x}$; that is, $l_{\boldsymbol{x},\boldsymbol{p}} \le u$ in Ω and $l_{\boldsymbol{x},\boldsymbol{p}}(\boldsymbol{x}) = u(\boldsymbol{x})$, where $l_{\boldsymbol{x},\boldsymbol{p}}$ is a supporting plane to u at $\boldsymbol{x}$.

The Brenier's potential u is differentiable at $\boldsymbol{x}$ if its subgradient $\partial u(\boldsymbol{x})$ is a singleton. We classify the points according to the dimensions of their subgradients, and define the sets $\Sigma_k(u) = \left\{ \boldsymbol{x} \in \mathbb{R}^d \,|\, \dim[\partial u(\boldsymbol{x})] = k \right\}$, $k = 0, 1, 2, \ldots, d$.

It is obvious that $\Sigma_0(u)$ is the set of regular points, and $\Sigma_k(u)$, where $k > 0$, is the set of singular points. We also define the reachable subgradients at $\boldsymbol{x}$ as follows:

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

It is well known that the subgradient is equal to the convex hull of the reachable subgradient:

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

Theorem 3.15 (regularity). Let $\Omega, \Lambda \subset \mathbb{R}^d$ be two bounded open sets, let $f, g : \mathbb{R}^d \to \mathbb{R}^+$ be two probability densities that are zero outside Ω and Λ, and are bounded away from zero and infinity on Ω and Λ, respectively. Denote by $T = \nabla u : \Omega \to \Lambda$ the OT map provided by Theorem 3.7. Then there exist two relatively closed sets $\Sigma_\Omega \subset \Omega$ and $\Sigma_\Lambda \subset \Lambda$ with $\Sigma_\Omega = \Sigma_\Lambda = 0$ such that T: $\Omega \backslash \Sigma_\Omega \to \Lambda \backslash \Sigma_\Lambda$ is a homeomorphism of class $C^{0,\alpha}_{\text{loc}}$ for some $\alpha > 0$.

We call Σ_Ω the singular set of the OT map $\nabla u : \Omega \to \Lambda$. Fig. 5 illustrates the singularity set structure, computed using the algorithm based on Theorem 4.2. We obtain the following:

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

The subgradient of $\boldsymbol{x}_0$, $\partial u(\boldsymbol{x}_0)$, is the entire inner hole of Λ, while $\partial u(\boldsymbol{x}_1)$ is the shaded triangle. For each point on $\gamma_k(t)$, $\partial u[\gamma_k(t)]$ is a line segment outside Λ. $\boldsymbol{x}_1$ is the bifurcation point of γ_1, γ_2, and γ_3. The Brenier's potential on Σ_1 and Σ_2 is not differentiable, and the OT map ∇u on them is discontinuous.

Journal Pre-proofs

Fig. 5. Singularity structure of an OT map.

4. Computational algorithm

Brenier's theorem can be directly generalized to the discrete situation. In GAN models, the source measure μ is given as a uniform (or Gaussian) distribution defined on a compact convex domain Ω; the target measure ν is represented as the empirical measure, which is the sum of the Dirac measures:

$$\nu=\sum_{i=1}^{n} v_i \delta\left(\boldsymbol{y}-\boldsymbol{y}_i\right) \quad (17)$$

where $Y = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_n\}$ are training samples, with the weights $\sum_{i=1}^{n} v_i = \mu(\Omega)$; δ is the characteristic function..

Each training sample $\boldsymbol{y}_i$ corresponds to a supporting plane of the Brenier's potential, denoted as follows:

$$\pi_{\boldsymbol{h},i}\left(\boldsymbol{x}\right)=\left\langle \boldsymbol{x}, \boldsymbol{y}_i\right\rangle + h_i \quad (18)$$

where the height h_i is an unknown variable. We represent all the height variables as $\boldsymbol{h} = (\boldsymbol{h}_1, \boldsymbol{h}_2,\ldots, \boldsymbol{h}_n)$.

An envelope of a family of hyper-planes in the Euclidean space is a hypersurface that is tangential to each member of the family at some point, and these points of tangency together form the whole envelope. As shown in Fig. 6, the Brenier's potential $u_{\boldsymbol{h}}$: $\Omega \to \mathbb{R}$ is a piecewise linear convex function determined by $\boldsymbol{h}$, which is the upper envelope of all its supporting planes:

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

Fig. 6. (a) Piecewise linear Brenier's potential ($u_{\boldsymbol{h}}$) and (b) its Legendre transformation $u_{\boldsymbol{h}}^*$. $\boldsymbol{\pi}_{\boldsymbol{h},i}^*$: the Legendre dual of $\pi_{\boldsymbol{h},i}$; $\nabla u_{\boldsymbol{h}}$: the gradient of $u_{\boldsymbol{h}}$; Proj: project map; Proj*: the projection map in the Legendre dual space.

The graph of the Brenier's potential is a convex polytope. Each supporting plane $\pi_{\boldsymbol{h},i}$ corresponds to a facet of the polytope. The projection of the polytope induces a cell decomposition of Ω, where each supporting plane $\pi_i(\boldsymbol{x})$ projects onto a cell $W_i(\boldsymbol{h})$, $\boldsymbol{p}$ is any point in $\mathbb{R}^d$:

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

The cell decomposition is a power diagram. The μ-measure of $W_i \cap \Omega$ is denoted as $w_i(\boldsymbol{h})$:

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

The gradient map $\nabla u_{\boldsymbol{h}} : \Omega \to Y$ maps each cell $W_i(\boldsymbol{h})$ to a single point $\boldsymbol{y}_i$:

$$\nabla u_{\boldsymbol{h}} : W_i\left(\boldsymbol{h}\right)\to \boldsymbol{y}_i,\ i=1,2,\ldots,n. \quad (22)$$

Given the target measure ν in Eq. (17), there exists a discrete Brenier's potential in Eq. (19) whose projected μ volume of each facet $w_i(\boldsymbol{h})$ is equal to the given target measure ν_i. This was proved by Alexandrov [46] in convex geometry.

Theorem 4.1 (Alexandrov [46]). Suppose Ω is a compact convex polytope with a non-empty interior in $\mathbb{R}^n$, $\boldsymbol{n}_1,\ldots,\boldsymbol{n}_k \subset \mathbb{R}^{n+1}$ are distinct k unit vectors, the $(n + 1)$th coordinates are negative, and $\nu_1, \ldots, \nu_k > 0$ so that $\sum_{i=1}^{k} v_i = \mathrm{vol}(\Omega)$. Then there exists a

Journal Pre-proofs

convex polytope $P \subset \mathbb{R}^{n+1}$ with the exact k codimension-1 faces $F_1,..., F_k$ so that $\boldsymbol{n}_i$ is the normal vector to F_i and the intersection between Ω and the projection of F_i has the volume v_i. Furthermore, such P is unique up to vertical translation.

Alexandrov's proof for the existence of the solution is based on algebraic topology, which is not constructive. Recently, Gu et al. [6] provided a constructive proof based on the variational approach.

Theorem 4.2 (Ref. [6]). Let μ be a probability measure defined on a compact convex domain Ω in $\mathbb{R}^d$, and let $Y = \{\boldsymbol{y}_1, \boldsymbol{y}_2,..., \boldsymbol{y}_n\}$ be a set of distinct points in $\mathbb{R}^d$. Then for any $v_1, v_2,..., v_n > 0$ with $\sum_{i=1}^{n} v_i = \mu(\Omega)$, there exists $\boldsymbol{h} = (\boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_n) \in \mathbb{R}^n$, which is unique up to adding a constant $(c, c,..., c)$, so that $w_i(\boldsymbol{h}) = v_i$, for all i. The vector $\boldsymbol{h}$ is the unique minimum argument of the following convex energy:

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

defined on an open convex set

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

Furthermore, $\nabla u_{\boldsymbol{h}}$ minimizes the quadratic cost

$$\frac{1}{2} \int_{\Omega} \left\| \boldsymbol{x} - T(\boldsymbol{x}) \right\|^2 \mathrm{d}\,\mu(\boldsymbol{x}) \quad (25)$$

among all transport maps $T_{\#}\mu = \nu$.

The gradient of the above convex energy in Eq. (23) is given by the following:

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

The ith row and jth column element of the Hessian of the energy is given by the following:

$$\frac{\partial w_i}{\partial h_j} = -\frac{\mu(W_i \cap W_j \cap \Omega)}{\left\| \boldsymbol{y}_i - \boldsymbol{y}_j \right\|}, \quad \frac{\partial w_i}{\partial h_i} = \sum_{j \neq i} \frac{\partial w_i}{\partial h_j} \quad (27)$$

As shown in Fig. 6, the Hessian matrix has an explicit geometric interpretation. Fig. 6(a) shows the discrete Brenier's potential $u_{\boldsymbol{h}}$, while Fig. 6(b) shows its Legendre transformation $u_{\boldsymbol{h}}^*$ using Definition 3.8. The Legendre transformation can be constructed geometrically: For each supporting plane $\pi_{\boldsymbol{h},i}$, we construct the dual point $\boldsymbol{\pi}_{\boldsymbol{h},i}^* = (\boldsymbol{y}_i, h_i)$; the convex hull of the dual points $\left\{ \boldsymbol{\pi}_{\boldsymbol{h},1}^*, \boldsymbol{\pi}_{\boldsymbol{h},2}^*, ..., \boldsymbol{\pi}_{\boldsymbol{h},n}^* \right\}$ is the graph of the Legendre transformation $u_{\boldsymbol{h}}^*$.

The projection of $u_{\boldsymbol{h}}^*$ induces a triangulation of $Y = \{\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_n\}$, which is the weighted Delaunay triangulation. As shown in Fig. 7, the power diagram in Eq. (20) and the weighted Delaunay triangulation are Poincaré dual to each other: If, in the power diagram, $W_i(\boldsymbol{h})$ and $W_j(\boldsymbol{h})$ intersect at a $(d-1)$-dimensional cell, then in the weighted Delaunay triangulation, $\boldsymbol{y}_i$ connects with $\boldsymbol{y}_j$. The element of the Hessian matrix in Eq. (27) is the ratio between the μ volume of the $(d-1)$ cell in the power diagram and the length of the dual edge in the weighted Delaunay triangulation.

Fig. 7. Power diagram (blue) and its dual-weighted Delaunay triangulation (black).

The conventional power diagram can be closely related to the above theorem.

Definition 4.3 (power distance). Given a point $\boldsymbol{y}_i \in \mathbb{R}^d$ with a power weight ψ_i, the power distance is given by the following:

Journal Pre-proofs

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

Definition 4.4 (power diagram). Given the weighted points $\left(\boldsymbol{y}_1,\psi_1\right),\ldots,\left(\boldsymbol{y}_k,\psi_k\right)$, the power diagram is the cell decomposition of $\mathbb{R}^d$:

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

where each cell is a convex polytope:

$$W_i\left(\psi\right)=\left\{\boldsymbol{x}\in\mathbb{R}^d\,\middle|\,\mathrm{pow}\left(\boldsymbol{x},\boldsymbol{y}_i\right)\le\mathrm{pow}\left(\boldsymbol{x},\boldsymbol{y}_j\right)\right\} \quad (30)$$

The weighted Delaunay triangulation, denoted as $T(\psi)$, is the Poincaré dual to the power diagram; if $W_i\left(\psi\right)\cap W_j\left(\psi\right)\neq\phi$, then there is an edge connecting $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ in the weighted Delaunay triangulation. Note that $\mathrm{pow}(\boldsymbol{x},\boldsymbol{y}_i)\le\mathrm{pow}(\boldsymbol{x},\boldsymbol{y}_j)$ is equivalent to

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

Let $h_i=1/2\left(\psi_i-\left\|\boldsymbol{y}_i\right\|^2\right)$; then we rewrite the definition of $W_i\left(\psi\right)$ as follows:

$$W_i\left(\psi\right)=\left\{\boldsymbol{x}\in\mathbb{R}^d\,\middle|\,\left\langle\boldsymbol{x},\boldsymbol{y}_i\right\rangle+h_i\ge\left\langle\boldsymbol{x},\boldsymbol{y}_j\right\rangle+h_j,\ \forall j\right\} \quad (32)$$

In practice, our goal is to compute the discrete Brenier's potential Eq. (19) by optimizing the convex energy Eq. (23). For low-dimensional cases, we can directly use Newton's method by computing the gradient Eq. (26) and the Hessian matrix Eq. (27). For deep learning applications, direct computation of the Hessian matrix is unfeasible; instead, we can use the gradient descend method or quasi-Newton's method with superlinear convergence. The key of the gradient is to estimate the μ volume $w_i(\boldsymbol{h})$. This can be done using the Monte Carlo method: We draw n random samples from the distribution μ, and count the number of samples falling within $W_i(\boldsymbol{h})$, which is the ratio converging to the μ volume. This method is purely parallel and can be implemented using a GPU. Moreover, we can use a hierarchical method to further improve the efficiency: First, we classify the target samples to clusters, and compute the OT map to the mass centers of the clusters; second, for each cluster, we compute the OT map from the corresponding cell to the original target samples within the cluster.

In order to avoid mode collapse, we need to find the singularity sets in Ω. As shown in Fig. 8, the target Dirac measure has two clusters; the source is the uniform distribution on the unit planar disk. The graph of the Brenier's potential function is a convex polyhedron with a ridge in the middle. The projection of the ridge on the disk is the singularity set $\Sigma_1(u)$, and the optimal mapping is discontinuous on Σ_1. In general cases, if two cells $W_i(\boldsymbol{h})$ and $W_j(\boldsymbol{h})$ are adjacent, then we compute the angle between the normals to the corresponding support planes:

$$\theta_{ij}=\frac{\left\langle\boldsymbol{y}_i,\boldsymbol{y}_j\right\rangle}{\left\|\boldsymbol{y}_i\right\|\left\|\boldsymbol{y}_j\right\|}$$

Fig. 8. Singularity set of the Brenier's potential function and discontinuity set of the OT map.

Journal Pre-proofs

If θ_{ij} is greater than a threshold, then the common facet $W_i(\boldsymbol{h}) \cap W_j(\boldsymbol{h})$ is in the discontinuity singular set.

5. GANs and optimal transportation

OT theory lays down the theoretical foundation for GANs. Many recent works, such as the WGAN [1], WGAN-GP [27], and RW-GAN [47], use the Wasserstein distance to measure the deviation between the generated data distribution and the real data distribution.

From the OT perspective, the optimal solutions for the generator and discriminator are related by a closed form; hence, the generator and the discriminator should collaborate instead of compete. More details can be found in Ref. [11]. Furthermore, the regularity theory of the solutions to the Monge–Ampère equation can explain the mode collapse in GANs [48].

5.1. Competition versus collaboration

The OT view of the WGAN [1] is illustrated in Fig. 2. According to the manifold distribution hypothesis, the real data distribution ν is close to a manifold Σ embedded in the ambient space χ. The generator computes the decoding map g_θ from the latent space Z to the ambient space, and transforms the white noise ζ (i.e., the Gaussian distribution) to the generated distribution, μ_θ. The discriminator computes the Wasserstein distance between μ_θ and the real data distribution ν, $W_c(\mu_\theta, \nu)$, by computing the Kantorovich's potential φ_ξ. Both g_θ and φ_ξ are realized by DNNs.

In the training process, the generator improves g_θ in order to better approximate ν by $((g_\theta)_\#\zeta$; the discriminator refines the Kantorovich's potential φ_ξ to improve the estimation of the Wasserstein distance. The generator and the discriminator compete with each other, without sharing intermediate computational results. Under the L^1 cost function, the alternative training process of the WGAN can be formulated as the min–max optimization of expectations:

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

But if we change the cost function to be L^2 distance, then according to Theorem 3.10, at the optimum, the Brenier's potential u and the Kontarovich's potential φ are related by a closed form of Eq. (16), $u(\boldsymbol{x}) = 1/2\|\boldsymbol{x}\|^2 - \varphi(\boldsymbol{x})$. The generator pursues the OT map ∇u; the discriminator computes φ. Hence, once the generator reaches the optimum, the optimal solution for the discriminator can be obtained without any training, and vice versa.

In more detail, suppose at the kth iteration, the generator map is g_θ^k. The discriminator computes the Kontarovich's potential φ_ξ, which gives the Wasserstein distance between the current generated data distribution $\left(g_\theta^k\right)_\# \zeta$ and the real data distribution ν; ∇u gives the OT map from $\left(g_\theta^k\right)_\# \zeta$ to ν. Therefore, we obtain the following:

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

This means that the generator map can be updated by the following:

$$g_\theta^{k+1} = \left(\mathrm{id} - \nabla\varphi_\xi\right)\circ g_\theta^k \qquad (33)$$

This conclusion shows that, in principle, the training process of the generator can be skipped; in practice, the efficiency can be improved greatly by sharing the intermediate computational results. Therefore, in designing the architectures of GANs, collaboration is better than competition.

5.2. Mode collapse and regularity

Although GANs are powerful for many applications, they have critical drawbacks: First, the training of GANs is tricky, sensitive to hyperparameters, and difficult to converge; second, GANs suffer from mode collapsing; and third, GANs may generate unrealistic samples. The difficulties of convergence, mode collapse, and the generation of unrealistic samples can be explained by the regularity Theorem 3.15 of the OT map.

According to Brenier's polar factorization, Theorem 3.9, any measure-preserving map can be decomposed into two maps, one of which is an OT map, which is a solution to the Monge–Ampère equation. According to the regularity Theorem 3.15, if the support Λ of the target measure ν has multiple connected components—that is, if ν has multiple modes, or Λ is non-convex—then the OT map T: $\Omega \to \Lambda$ is discontinuous on the singular set Σ_Ω.

Fig. 9 shows the multi-cluster case: Λ has two connected components, where the OT map T is discontinuous along Σ_1. Fig. 10 shows that even Λ is connected, albeit non-convex. Ω is a rectangle, Λ is a dumbbell shape, the density functions are constants, the OT map is discontinuous, and the singularity set $\Sigma_1 = \gamma_1 \bigcup \gamma_2$.

Journal Pre-proofs

(a) (b)

Fig. 9. Discontinuous OT map, produced by a GPU implementation of an algorithm based on Theorem 4.2. (a) is the source domain and (b) is the target domain. The middle line in (a) is the singularity set Σ_1.

(a) (b)

Fig. 10. Discontinuous OT map, produced by a GPU implementation of an algorithm based on Theorem 4.2. (a) is the source domain and (b) is the target domain. γ_1 and γ_2 in (a) are two singularity sets.

Fig. 11 shows an OT map between two probability measures in $\mathbb{R}^3$. Both the source measure μ and the target measure ν are uniform distributions; the support of Ω is the unit solid ball, and the support of Λ is the solid Stanford bunny. We compute the Brenier's potential u: $\Omega \to \mathbb{R}$ based on Theorem 4.2. In order to visualize the mapping, we interpolate the probability measure as follows:

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

Fig. 11 shows the support of the interpolated measure ρ_t. The foldings on the surface are the singularity sets, where the OT map is discontinuous.

(a) (b) (c) (d)

Fig. 11. OT from the Stanford bunny to a solid ball. The singular sets are the foldings on the boundary surface. (a–d) shows the deformation procedure.

In a general situation, due to the complexity of the real data distributions, the embedding manifold Σ, and the encoding/decoding maps, the supports of the target measures are rarely convex; therefore, the transportation mapping cannot be globally continuous.

On the other hand, general DNNs, such as ReLU DNNs, can only approximate continuous mappings. The functional space represented by ReLU DNNs does not contain the desired discontinuous transportation mapping. The training process or, equivalently, the searching process will lead to three alternative situations:

(1) The training process is unstable, and does not converge.

Journal Pre-proofs

(2) The searching converges to one of the multiple connected components of Λ, and the mapping converges to one continuous branch of the desired transportation mapping. This means that a mode collapse is encountered.

(3) The training process leads to a transportation map, which covers all the modes successfully, but also covers the regions outside Λ. In practice, this will induce the phenomenon of generating unrealistic samples, as shown in the middle frame of Fig. 12.

Therefore, in theory, it is impossible to approximate OT maps directly using DNNs.

5.3. AE–OT model

As shown in Fig. 4, we separate the two main tasks of GANs: manifold learning and probability distribution transformation. The first task is carried out by an AE to compute the encoding/decoding maps f_θ, g_ξ; the second task is accomplished using the explicit variational method to compute the OT map T in the latent space. The real data distribution ν is pushed forward by the encoding map f_θ , inducing $(f_\theta)_\#\nu$. In the latent space, T maps the uniform distribution μ to $(f_\theta)_\#\nu$.

The AE–OT model has many advantages. In essence, finding the OT map is a convex optimization problem; the existence and the uniqueness of the solution are guaranteed. The training process is stable and has superlinear convergence by using quasi-Newton's method. The number of unknowns is equal to that of the training samples, avoiding over-paramerization. The parallel OT map algorithm can be implemented using a GPU. The error bound of the OT map can be controlled by the sampling density in the Monte Carlo method. The hierarchical algorithm with self-adaptivity further improves the efficiency. In particular, the AE–OT model can eliminate mode collapse.

6. Experimental results

In this section, we report our experimental results.

6.1. Training process

The training of the AE–OT model mainly includes two steps: training the AE and finding the OT map. The OT step is accomplished using a GPU implementation of the algorithm, as described in Section 4. In the AE step, during the training process, we adopt the Adam algorithm [49] to optimize the parameters of the neutral network, with a learning rate of 0.003, $\beta_1 = 0.5$, and $\beta_2 = 0.999$. When the L^2 loss stops descending, which means that the network has found a good encoding map, we freeze the encoder part and continue to train the network for the decoding map. The training loss before and after the freezing of the encoder is shown in Table 1. Next, in order to find the OT map from the given distribution (here, we use uniform distribution) to the distribution of latent features, we randomly sample $100N$ random points from the uniform distribution to compute the gradient of the energy. Here, N is the number of latent features of the dataset. Also, in the experiment, θ_{ij} is set to be different for different datasets. To be specific, for the MNIST and Fashion-MNIST datasets, θ_{ij} is set to be 0.75, while for the CIFAR-10 and CelebA datasets, it is set to be 0.68 and 0.75, respectively.

Table 1
The L^2 loss of the AEs before and after the freezing of the encoder.

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

Our AE–OT model was implemented using PyTorch on a Linux platform. All the experiments were conducted on a GTX1080Ti.

6.2. Transportation map discontinuity test

In this experiment, we want to test our hypothesis: In most real applications, the support of the target measure is non-convex, the singularity set is non-empty, and the probability distribution map is discontinuous along the singularity set.

As shown in Fig. 12, we use an AE to compute the encoding/decoding maps from the CelebA dataset (Σ, ν) to the latent space Z; the encoding map $f_\theta : \Sigma \to Z$ pushes forward ν to $(f_\theta)_\#\nu$ on the latent space. In the latent space, we compute the OT map based on the algorithm described in Section 4, T: $Z \to Z$, where T maps the uniform distribution in a unit cube ζ to $(f_\theta)_\#\nu$. Then we randomly draw a sample z from the distribution ζ and use the decoding map g_ξ: $Z \to \Sigma$ to map $T(z)$ to a generated human facial image $g_\xi \circ T(z)$. Fig. 12(a) demonstrates the realistic facial images generated by this AE–OT framework.

Journal Pre-proofs

(a) (b)

Fig. 12. Facial images generated by an AE–OT model. (a) Generated realistic facial images; (b) a path through a singularity. The image in the center of (b) shows that the transportation map is discontinuous.

If the support of the push-forward measure $(f_\theta)_\# \nu$ in the latent space is non-convex, there will be a singularity set Σ_k, where $k > 0$. We would like to detect the existence of Σ_k. We randomly draw line segments in the unit cube in the latent space, and then densely interpolate along this line segment to generate facial images. As shown in Fig. 12(b), we find a line segment γ, and generate a morphing sequence between a boy with a pair of brown eyes and a girl with a pair of blue eyes. In the middle, we generate a face with one blue eye and one brown eye, which is definitely unrealistic and outside Σ. This result means that the line segment γ goes through a singularity set Σ_k, where the transportation map T is discontinuous. This also shows that our hypothesis is correct: The support of the encoded human facial image measure on the latent space is non-convex.

As a byproduct, we find that this AE–OT framework improves the training speed by a factor of five and increases the convergence stability, since the OT step is a convex optimization. Thus, it provides a promising way to improve existing GANs.

6.3. Mode collapse comparison

Since the synthetic dataset consists of explicit distributions and known modes, mode collapse can be accurately measured. We chose two synthetic datasets that have been studied or proposed in prior works [50,51]: a 2D grid dataset.

For a choice of the measurement metric of mode collapse, we adopted three previously used metrics [50,51]. Number of modes counts the quantity of modes captured by the samples produced by a generative model. In this metric, a mode is considered as lost if no sample is generated within three standard deviations of that mode. Percentage of high-quality samples measures the proportion of samples that are generated within three standard deviations of the nearest mode. The third metric, used in Ref. [51], is the reverse Kullback–Leibler (KL) divergence. In this metric, each generated sample is assigned to its nearest mode, and we count the histogram of samples assigned on each mode. This histogram then forms a discrete distribution, whose KL divergence with the histogram formed by real data is then calculated. Intuitively, this measures how well the generated samples balance among all modes regarding the real distribution.

In Ref. [51], the authors evaluated GAN [26], ALI [52], MD [30], and PacGAN [51] on synthetic datasets with the above three metrics. Each experiment was trained under the same generator architecture with a total of approximately 400k training parameters. The networks were trained on 100k samples for 400 epochs. For the AE–OT experiment, since the source space and target space are both 2D, there is no need to train an AE. We directly compute a semi-discrete OT that maps between the uniform distribution on the unit square and the empirical real data distribution. Theoretically, the minimum amount of real sample needed for OT to recover all modes is one sample per mode. However, this may lead to the generation of low-quality samples during the interpolation process. Therefore, for OT computation, we take 512 real samples, and new samples are generated based on this map. We note that, in this case, there are only 512 parameters to optimize in OT computing, and the optimization process is stable due to the existence of the convex positive-definite Hessian. Our results are provided in Table 2, and benchmarks of previous methods are copied from Ref. [51]. For illustration purposes, we plotted our results on synthetic datasets along with those of GAN and PacGAN in Fig. 13.

Table 2
Mode collapse comparison for the 2D grid dataset.

	Modes	Samples	Reverse KL
GAN	17.3 ± 0.8	94.8 ± 0.7%	0.70 ± 0.07

Journal Pre-proofs

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

Fig. 13. Mode collapse comparison on a 2D grid dataset. (a) GAN; (b) PacGAN4; (c) AE–OT. Orange marks are real samples and green marks are generated ones.

6.4. Comparison with the state of the art

We designed experiments to compare our proposed AE–OT model with state-of-the-art generative models, including the adversarial models evaluated by Lucic et al. in Ref. [33], and the non-adversarial models studied by Hoshen and Malik in Ref. [36].

For the purpose of fair comparison, we used the same testing datasets and network architecture. The datasets included MNIST [53], MNIST-Fashion [54], CIFAR-10 [55], and CelebA [56], similar to those tested in Refs. [31,36]. The network architecture was similar to that used by Lucic et al. in Ref. [33]. In particular, in our AE–OT model, the network architecture of the decoder was the same as that of the generators of GANs in Ref. [33], and the encoder was symmetric to the decoder.

We compared our model with state-of-the-art generative models using the FID score [31] and PRD curve as the evaluation criteria. The FID score measures the visual fidelity of the generated results and is robust to image corruption. However, the FID score is sensitive to mode addition and dropping [33]. Hence, we also used the PRD curve, which can quantify the degree of mode dropping and mode inventing on real datasets [32].

6.4.1. Comparison with FID score

The FID score is computed as follows: ① Extract the visually meaningful features of both the generated and real images by running the inception network [30]; ② fit the real and generated feature distributions with Gaussian distributions; and ③ compute the distance between the two Gaussian distributions using the following formula:

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

where μ_r and μ_g represent the means of the real and generated distributions, respectively; and Σ_r and Σ_g represent the variances of these distributions.

The comparison results are summarized in Tables 3 and 4. The statistics of various GANs come from Lucic et al. [33], and those of the non-adversarial generative models come from Hoshen and Malik [36]. In general, our proposed model achieves better FID scores than the other state-of-the-art generative models.

Table 3
Quantitative comparison with FID-I.

Dataset	Adversarial				
	MM GAN	NS GAN	LSGAN	WGAN	BEGAN
MNIST	9.8	6.8	7.8	6.7	13.1
Fashion-MNIST	29.6	26.5	30.7	21.5	22.9

Journal Pre-proofs

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.

Table 4
Quantitative comparison with FID-II.

Dataset	Non-adversarial			Reference	
	VAE	GLO	GLANN	AE	AE–OT
MNIST	23.8	49.6	8.6	5.5	**6.4**
Fashion-MNIST	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.

Theoretically, the FID scores of our AE–OT model should be close to those of the pre-trained AEs; this is also validated by our experiments.

The fixed network architecture of our AE was adopted from Lucic et al. [33]; its capacity is not large enough to encode CIFAR-10 or CelebaA, so we had to down-sample these datasets. We randomly selected 25k images from CIFAR-10 and 10k images from CelebA to train our model. Even so, our model obtained the best FID score in CIFAR-10. Dut to the limited capacity of the InfoGAN model, the performance of the AE of CelebA, whose FID of 67.5 is not ideal, further caused the FID of the generated dataset to be 68.4. By adding two more convolutional layers to the AE architecture, the L^2 loss in CelebA was less than 0.03, and the FID score beat all other models (28.6, as shown in the bracket of Table 4).

6.4.2. Comparison with the PRD curve

The FID score is an effective method to measure the difference between the generated distribution and the real data distribution, but it mainly focuses on precision, and cannot accurately capture what portion of real data a generative model can cover. The method proposed in Ref. [32] disentangles the divergence between distributions into two components: precision and recall.

Given a reference distribution P and a learned distribution Q, the precision intuitively measures the quality of samples from Q, while the recall measures the proportion of P that is covered by Q.

We used the concept of (F_8, $F_{1/8}$) introduced by Sajjadi et al. in Ref. [32] to quantify the relative importance of precision and recall. Fig. 14 summarizes the comparison results. Each dot represents a specific model with a set of hyperparameters. The closer a dot is to the upper-right corner, the better the performance of the model is. The blue and green dots show the GANs and VAEs evaluated in Ref. [32], the yellow dot represents the GLANN model in Ref. [36], and the red dot is our AE–OT model.

It is clear that our proposed model outperforms others for MNIST and Fashion-MNIST. For the CIFAR-10 dataset, the precision of our model is slightly lower than those of GANs and GLANN, but the recall is the highest. For the CelebA dataset, due to the limited capacity of the AE, the performance of our model is not impressive. However, after adding two more convolutional layers to the AE, our model achieves the best score.

Journal Pre-proofs

Journal Pre-proofs

Fig. 14. A comparison of the precision-recall pair in (F_8, $F_{1/8}$) in the four datasets. (a) MNIST; (b) Fashion-MNIST; (c) CIFAR-10; (d) CelebA. The khaki dots are the results of Ref. [36]. The red dots are the results of the proposed method. The purple dot in the fourth subfigure corresponds to the results of the architecture with two more convolutional layers.

6.4.3. Visual comparison

Fig. 15 shows a visual comparison between the images generated by our proposed method and those generated by the GANs studied by Lucic et al. in Ref. [33] and the non-adversarial models studied by Hoshen and Malik in Ref. [36]. The first column shows the original images, the second column shows the results generated by the AE, the third column shows the best generating results of the GANs in Lucic et al. [33], the fourth column displays the results generated by the models of Hoshen and Malik [36], and the fifth column displays the results from our method. It is clear that our method generates high-quality images and covers all modes.

Journal Pre-proofs

Journal Pre-proofs

Fig. 15. A visual comparison of the four datasets. The first column (a) shows the real data; the second column (b) is generated by an AE; the third column (c) illustrates the generating results of the GANs [33] with the highest precision-recall scores of (F_8, $F_{1/8}$), corresponding to the B dots in Fig. 14; the fourth column (d) gives the results of Ref. [36]; and the last column (e) shows the results of the proposed method.

7. Conclusion

This work uses OT theory to interpret GANs. According to the data manifold distribution hypothesis, GANs mainly accomplish two tasks: manifold learning and probability distribution transformation. The latter task can be carried out using the OT method directly. This theoretical understanding explains the fundamental reason for mode collapse, and shows that the intrinsic relation between the generator and the discriminator should be collaboration instead of competition. Furthermore, we propose an AE–OT model, which improves the theoretical rigor, training stability, and efficiency, and eliminates mode collapse.

Our experiment validates our assumption that if the distribution transportation map is discontinuous, then the existence of the singularity set leads to mode collapse. Furthermore, when our proposed model is compared with the state of the art, our method eliminates the mode collapse and outperforms the other models in terms of the FID score and PRD curve.

In the future, we will explore the theoretical understanding of the manifold learning stage, and use a rigorous method to make this part of the black box transparent.

Acknowledgements

The project is partially supported by the National Natural Science Foundation of China (61936002, 61772105, 61432003, 61720106005, 61772379), NSF CMMI-1762287 collaborative research: computational framework for designing conformal stretchable electronics, Ford URP topology optimization of cellular mesostructures' nonlinear behaviors for crash safety, and NSF DMS-1737812 collaborative research: ATD: theory and algorithms for discrete curvatures on network data from human mobility and monitoring.

Reference

[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.

Journal Pre-proofs

[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.

Journal Pre-proofs

(a) (b)

Journal Pre-proofs

(a) (b)

(a) (b)

Journal Pre-proofs

(a) (b)

(a) (b) (c) (d)

(a) (b)

(a) (b) (c)

Journal Pre-proofs

Journal Pre-proofs

Journal Pre-proofs

Journal Pre-proofs

(a) (b) (c) (d) (e)

Journal Pre-proofs