生成对抗网络的几何观点以及算法优化

项目研究报告

摘　要

生成对抗网络（GAN）模型作为一个生成模型，在近些年很受欢迎。它可以在不知道目标解析表达式的情况下给出一个从简单分布(例如高斯分布) 到目标分布的映射。生成对抗网络能够轻易地学习并生成图像等高维对象。

为了解决在对抗生成网络（GAN）中出现的梯度消失问题，人们提出了许多方法，引人关注的 WGAN 就是其中之一。它采用了最优传输理论中的 Wasserstein 距离作为判别器的损失函数。通过最优传输中的 Kantorovich 方法和凸几何中的 Brenier 方法，我们可以对 WGAN 进行解释，并证明特定情况下所有 GAN 类模型的对抗训练是不必要的。

最后，我们基于变分自编码器和最优传输理论提出一个新的生成模型 OTVAE。我们将其与 WGAN 进行对比，说明其比 WGAN 训练速度快，且生成的图片让人眼感觉更加贴近实际。

关键词：生成对抗网络，生成模型，微分流形，最优传输理论

Abstract

The Generative Adversarial Network (GAN) model has become popular in recent years. It can give a mapping from a simple distribution (e.g., Gaussian distribution) to the target distribution without knowing its explicit expression. Generative adversarial networks can easily learn and generate high-dimensional objects such as image.

To solve the gradient vanishing problem that arises in GANs, many approaches have been proposed, and the intriguing WGAN is one of them. It adopts the Wasserstein distance from optimal transportation theory as the loss function of the discriminator. With the Kantorovich's approach in optimal transport and the Brenier's approach in convex geometry, we can interpret WGAN and prove that adversarial training of all GAN-like models in some cases is unnecessary.

In the end, we propose a new generative model OTVAE based on variational auto-encoder and optimal transmission theory. We compare it with WGAN and show that it is faster to train than WGAN and generates images that feel more realistic to the human eye.

Key Words: Generative Adversarial Networks; Generative Models; Differentiable Manifold; Optimal Transportation Theory

目录

目录

第一章　简介

生成式对抗网络 (Generative Adversarial Networks，简称 GAN)[10]，作为一种无监督学习，它已然成为机器学习领域中一颗冉冉升起的新星。它可以生成与样本极其相似甚至可以以假乱真的文字、图像、视频等数据，让我们得以一瞥强人工智能的未来。除此之外，GAN 有着非常有趣的训练模式：生成器提高生成能力来生成更逼真的数据，判别器提高鉴别能力，它们相互竞争，最终形成均衡，如果运气好的话两者最后都可以达成理想的训练结果。

但是结果往往并不如意，除了优化速度较慢之外，GAN 还存在着无法解释、梯度消失[11]、模式崩溃[33][34][35] 等问题。针对 GAN 的缺点，有很多学者提出了许多的改进方法，例如深度卷积生成对抗网络 (DCGAN)[12]，Wasserstein GAN[11]，WGAN 的改进版 WGAN-GP[37] 等等。这些方法对 GAN 的一些模棱两可的技巧进行了改进，有的替换了损失函数的评价方法，在解决梯度消失和模式崩溃等问题上取得了不错的成果。但是他们都没有解决 GAN 的黑箱问题，我们对 GAN 进行学习和生成的原理尚不清楚。而顾险峰教授等人在 [4] 中提出的从几何观点审视生成对抗网络的观点令我们醍醐灌顶。在这个观点下，一些常见情况下，生成对抗网络中的对抗过程是没有必要的，在 L^2 代价函数下，生成器和判别器的训练是相互促进的。

顾教授等人虽然提出了 GAN 的几何解释的理论框架，但是在细节方面还不够明确；并且他没有明确给出 GAN 的改进算法。我们在这里利用微分流形和最优传输的知识解释 GAN 以及其衍生品的训练目标和过程，提高了 GAN 和生成模型的可解释性。我们还针对生成模型的训练过程进行优化，利用变分自编码器模型 (VAE)[38]，提出了一种新的生成模型 OTVAE。其与业界普遍认为性能优秀的 WGAN 比起来，在训练速度与生成结果质量上都取得了进步。

本文从 GAN 与 WGAN 的主要原理与结构开始，逐一对我们所用到的微分流形理论、最优传输理论进行介绍。然后介绍从几何角度和最优传输角度对 GAN 以及其他深度学习模型的解释。最后陈述离散最优传输问题的解法，介绍我们基于该解法的生成模型 OTVAE，并附上了与 WGAN 进行比较的结果。

第二章　生成对抗网络 (GAN)

第一节　GAN 的模型与原理

在最开始，作为我们算法的起点，我们首先介绍一下 GAN 的基本原理和模型。GAN[10] 由两部分组成：生成器 (Generator)G 和判别器 (Discriminator)D。

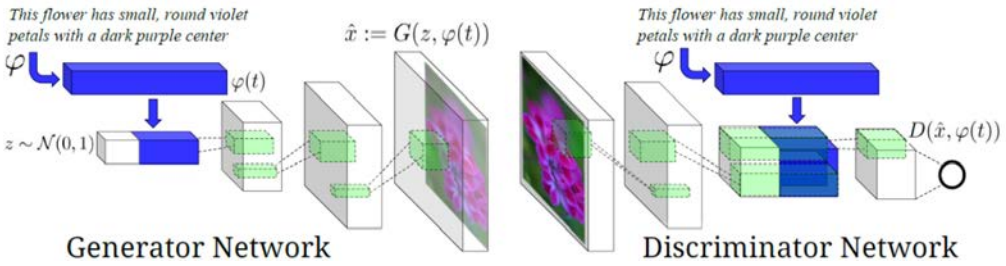

图 1: GAN 的模型[27]

从数学上来讲，生成器 G 是一个从给定的简单分布 (如高斯分布) 到目标分布 (如人脸图片的分布) 的映射。生成器接收我们提供的一个随机向量，生成一个样本。在本例子中生成器由深度神经网络来实现。判别器 D 是一个度量，用 JS 散度[29] 给出两个分布之间的“距离”(严格来说，JS 散度并不是一种距离，因为其不满足对称性)，借此评价每个样本是来自生成器还是训练集。这里判别器也由深度神经网络来实现。

生成器 G 和判别器 D 是同时相互进行训练的，在训练过程中我们想提高生成器生成真样本的能力，也想提高判别器判断正确的概率。这样看来对生成器和判别器的训练构成了一种对抗。学习的结果是对抗最后达到均衡，也就是到判别器无法分出样本是来自生成分布还是训练集分布为止。

原作者[10] 在设计 GAN 时，认为两者都使用多层感知机 (multilayer perceptron)[30] 时，对抗框架最容易实现。其提出了如下的对抗框架：

- 生成器 G：为了从训练集 X_{train} 中学习生成器的分布 p_g，定义一个输入噪声变量 $p_z(z)$，把从它到数据空间的映射表示为 $G(z;\Theta_g)$。这里的 G 是一个由参数为 Θ_g 的可微函数，由一个多层感知机实现。

- 判别器 D：第二个多层感知机 $D(x;\Theta_d)$，它输出一个参数为 Θ_d 的标量。$D(x)$ 表示样本 x 来自训练集分布 p_{data} 而不是生成分布 p_g 的概率。

- D 和 G 的目标：我们训练 D，使得它为训练集样本和 G 生成的样本分配正确标签的可能性最大化，同时训练 G 来最小化生成样本分布和给定分布之间的距离，即在给定 $D(x)$ 时最小化 $log(1 - D(G(z)))$ 的期望。

从优化的角度来说，D、G 进行下式的 min-max 博弈[10]：

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))] \tag{2.1}$$

我们在这里给出 GAN 的训练算法。其中超参数 m 为每一批学习的样本个数，超参数 k 指更新 k 步 G 后更新一次 D，算法中的随机梯度上升/下降可以换成任何其他的基于梯度的学习方法。

Algorithm 1 GAN 的小批量随机梯度下降训练

for 训练迭代次数 **do**

for k 次 **do**

- 从噪声先验 $p_g(z)$ 选取 m 个样本 $z^{(1)}, z^{(2)}, ..., z^{(m)}$
- 从生成器的生成分布 $p_{data}(x)$ 选取 m 个样本 $x^{(1)}, x^{(2)}, ..., x^{(m)}$
- 通过随机梯度上升更新判别器：

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log D\left(x^{(i)}\right) + \log\left(1 - D\left(G\left(z^{(i)}\right)\right)\right)\right] \tag{2.2}$$

end for

- 从噪声先验 $p_g(z)$ 选取 m 个小批量样本 $z^{(1)}, z^{(2)}, ..., z^{(m)}$
- 通过随机梯度下降更新生成器：

$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log\left(1 - D\left(G\left(z^{(i)}\right)\right)\right) \tag{2.3}$$

end for

原作者还证明了[10] 如果 G，D 有足够的容量，而且在算法的每个步骤中，判别器都达到了在给定的 G 下的最佳值，并且 p_g 也被更新从而满足公式 (2.1)，则 p_g 最终一定会收敛到 p_{data}。因此作者认为应将 $p_g = p_{data}$ 作为训练结束的条件。但是在实际应用中，不像其他的深度学习问题有严谨的性能度量，生成式模型只能靠肉眼观察训练的结果质量如何 (如主观判断生成的图片是否较为“真实”)，没有合理的判据来判断 $p_g = p_{data}$ 是否成立，这也是 GAN 的缺陷之一。虽然如此，在大部分情况下，训练次数足够多时，GAN 能实现能令人满意的效果。

第二节　GAN 的优势与缺陷

2.2.1　GAN 的优势

- 当真实数据的概率分布不可或难以计算的时候 (尤其对于图片、视频、语音等超高维数据，难以定义其上的经验分布)，传统依赖于数据内在解释的生成模型无法直接应用，但是生成对抗网络模型依然可以使用。
- 机器学习需要大量数据样本，而生成对抗网络模型自身理论上可以生成无尽的样本，从而降低了对于训练数据的需求，因此极其适合无监督学习。
- 理论上该框架可以训练所有的生成模型。

GAN 现在已经应用在许多领域，主要用途是用来生成常规情况下无法获得的大量样本，以及图像的修复和分辨率提高。例如 [31] 中提到的生成现实图片和 [32] 中提出的图片修复功能，都有很好的成果。

2.2.2　GAN 的缺点

- 如第一节所述，GAN 缺乏一个合理的判据来判断训练结果到底如何，只能通过人类凭感觉判断，当我们训练的目标不是图片这种易于判断的对象时，难以得知 $p_g = p_{data}$ 是否成立。
- GAN 的训练过程是不可视的，也就是俗称的“黑箱算法”。这导致 GAN 是一个很难解释的模型，导致我们很难解释其生成结果为什么好或为什么坏。近年来，一些学者从几何和最优传输理论方面对 GAN 进行了一些解释[4][5][33]，详见第五章。
- GAN 存在梯度消失问题。简单地说，判别器训练得越好，梯度就越趋向于 0，越容易出现梯度消失情况。此时优化就无从谈起。GAN 的优化模型 WGAN[11] 解决了这个问题，我们将在本章的第三节介绍 WGAN。
- GAN 存在模式崩溃[33][34][35] 的问题。比如当训练的图片种类只有猫一种时，可以得到不错的效果；但是如果同时混入了猫狗两种训练样本，此时 GAN 的生成结果就并不好。此外还存在一些现象：有时虽然涵盖了训练的所有模式，但是生成的样本有一些是与训练样本完全无关的，无意义的。这种问题在视频生成[39] 时尤其明显，GAN 会经常输出一些画面变换自然但实际内容无法令人理解的视频。而且 GAN 的生成结果非常随机，难以完全复现。该问题是 GAN 及其衍生品的最大通病。

第三节　WGAN 对 GAN 的改进

针对 GAN 存在的诸多问题，有许多的研究成果做出了一定的优化。这里重点介绍 Wasserstein GAN(简称为 WGAN)[11]，其引入了最优传输理论中的 Wasserstein 距离[14]。WGAN 采用 Wasserstein 距离作为判别器的损失函数，这样当生成分布和训练集样本分布的支撑集没有相交部分时，WGAN 就可以得到一个用来优化生成器的合适梯度。WGAN 作者进行了非常繁复的数学推导，这里略去证明，只对算法的内容和一些结论进行介绍。

先给出 WGAN 的小批量随机梯度下降训练的算法伪代码：

算法中的超参数如下：学习速率 α，裁剪参数 c，每批数据个数 m，每次生成器迭代时的判别器的迭代次数 k，初始判别器参数 ω_0，初始生成器参数 θ_0。

$RMSProp$ 指的是采用 RMSProp 算法的优化器[36]。

$clip(w,-c,c)$ 被称为权重裁剪 (weight clipping)，指的是将权重 w 裁剪至闭区间 $[-c,c]$ 内，这样能保证 g_w 是 Lipschitz 连续的。这样 Wasserstein 距离处处连续，几乎处处可微，从而避免梯度消失问题[11]。

Algorithm 2 WGAN 的小批量随机梯度下降训练

while θ 尚未收敛 **do**

for k 次 **do**

- 从真实数据分布 $\mathbb{P}_r$ 选取样本 $\{x^{(i)}\}_{i=1}^{m}$
- 从给定先验分布 $p(z)$ 选取样本 $\{z^{(i)}\}_{i=1}^{m}$
- $g_w \leftarrow \nabla_w \left[\frac{1}{m}\sum_{i=1}^{m}\int_w\left(x^{(i)}\right)-\frac{1}{m}\sum_{i=1}^{m}\int_w\left(g_\theta\left(z^{(i)}\right)\right)\right]$
- $w \leftarrow w+\alpha\cdot \text{RMSProp}\left(w,g_w\right)$
- $w \leftarrow \text{clip}(w,-c,c)$

end for

- 从给定先验分布 $p(z)$ 选取样本 $\{z^{(i)}\}_{i=1}^{m}$
- $g_\theta \leftarrow -\nabla_\theta\frac{1}{m}\sum_{i=1}^{m}\int_w\left(g_\theta\left(z^{(i)}\right)\right)$
- $\theta \leftarrow \theta-\alpha\cdot RMSProp\left(\theta,g_\theta\right)$

end while

与 GAN 相比 WGAN 所做的改动主要是三点：

- 生成器与判别器的损失函数没取对数.
- 判别器最后一层没有使用 sigmoid 函数.
- 更新判别器参数后进行权重裁剪.

WGAN 有以下优缺点：

- 得到了与生成器的收敛性以及生成样本质量相关的有意义的度量。在算法中的损失函数是一个对 Wasserstein 距离的估计，其趋向于一个与设置的 Lipschitz 常数有关的常数。但是这个估计也是模糊的，缺少显式的计算方法。并且通过权重裁剪保证 Lipschitz 连续是一个粗糙的想法，在 [37] 中对其进行了分析和改进。虽然如此，这仍然是对 GAN 的一个巨大改进，解决了 GAN 的梯度消失的问题。

- 提高了优化过程的稳定性。作者在众多实验中均未观测到模式崩溃现象，但无法证明模式崩溃问题已经解决。事实上，其并没有完全解决模式崩溃问题，在 [37] 中对其有详细分析。

第四节　WGAN 的后续改进

WGAN 也存在一定缺点，比如它训练困难，收敛速度慢。有一些人对 WGAN 提出了改进，比如蒙特利尔大学提出的 WGAN-GP 模型[37]。

他们发现 WGAN 仍然存在着模式崩溃现象，而原因通常是由于在 WGAN 中使用了权重裁剪来在判别器函数上实现 Lipschitz 约束。他们提出了一种替代权重裁剪的办法：根据判别器的输入来惩罚判别器的梯度的范数 (即 Gradient Penalty，简称为 GP)。这种方法比 WGAN 的性能更好，并且能在不调参的前提下对多种 GAN 架构进行训练。

他们认为通过权重裁剪这种方法会使判别器分布偏向更简单的函数，从而无法捕捉数据分布的高阶细节。他们提出了另外一种方法实施 Lipschitz 约束：由于一个可微函数是 1-Lipschtiz 的，当且仅当它在任何地方有至多为 1 的梯度范数，所以他们考虑针对判别器输入直接限制判别器输出的梯度范数。在实际处理中，他们对随机样本 $\widehat{x} \sim \mathbb{P}_{\widehat{x}}$ 放松了约束条件。他们使用的新的判别器函数是

$$L = \underbrace{\mathop{\mathbb{E}}_{\tilde{\boldsymbol{x}}\sim\mathbb{P}_g}[D(\tilde{\boldsymbol{x}})] - \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathbb{P}_r}[D(\boldsymbol{x})]}_{\text{WGAN 的损失函数}} + \underbrace{\lambda \mathop{\mathbb{E}}_{\hat{\boldsymbol{x}}\sim\mathbb{P}_{\hat{\alpha}}}\left[\left(\|\nabla_{\hat{\boldsymbol{x}}} D(\hat{\boldsymbol{x}})\|_2 - 1\right)^2\right]}_{\text{新增的梯度惩罚项}}. \tag{2.4}$$

在后面的实验中，研究者均使用了四种不同的 GAN 进行训练：WGAN-GP，WGAN，DCGAN，最小二乘 GAN，其中 WGAN-GP 是唯一一种使用同一种默认超参数，并在每个架构下都成功训练的方法。因此 WGAN-GP 要比 WGAN 更稳定，但原文并没有说明其彻底解决了模式崩溃的问题。

第三章　微分流形理论与最优传输理论简介

WGAN 和 WGAN-GP 虽然比 GAN 更加稳定，但他们并没有对 GAN 的学习及生成过程进行解释，我们依旧不知道 GAN 的内部机理。为此我们将视线投向纯数学方面，[4][5] 尝试利用微分几何和最优传输理论对 GAN 进行解释。

微分流形理论是微分几何学的重要组成部分，而它则是现在机器学习理论界流形的理论“流形假设”的基础，借此理论我们得以在抽象的数学理论与机器学习的黑箱之间建立起联系。而 WGAN 所采用的关键优化:Wasserstein 距离则来自最优传输理论。我们将在这一章对微分流形理论以及最优传输理论的基本概念以及一些必要的符号进行介绍，为后续对 WGAN 进行数学解释打下基础。

第一节　流形的概念与符号

我们在这里先给出微分流形中几个重要的概念，并对接下来要使用的符号进行定义：

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

$$\varphi_{\alpha\beta} := \varphi_\beta \circ \varphi_\alpha^{-1} \tag{3.1}$$

Definition 2. $Polish$ 空间[14][15][16][17]：完备，可分的度量空间 (X, d)。

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

Definition 4. 前推映射 *(push-forward mapping)*T 与前推算子 *(push-forward operator)*T^**(*或记为 $T_\#$*)*[14]：假设 X, Y 为两个 $Polish$ 空间，$T : X \to Y$ 是一个 $Borel$ 映射，$\mu \in \mathcal{P}(X)$ 是一个测度，则可以定义 Y 上的测度 $\nu = T^*\mu$：

$$T^*\mu(E) = \mu(T^{-1}(E)), \quad \forall E \subset Y,\ E\ Borel \tag{3.2}$$

我们称 T 将 μ 前推到了 ν。

设 $\mathcal{X}$ 是环境空间 (ambient space)，其是一个 Polish 空间，μ 是定义在 $\mathcal{X}$ 上的概率分布，表示为密度函数 $\mu : \mathcal{X} \to \mathbb{R}_{\geq 0}$。则 μ 的支撑集 $\Sigma(\mu) := supp(\mu) = \{\mathrm{x} \in \mathcal{X} | \mu(x) > 0\}$ 是 $\mathcal{X}$ 上的一个低维流形。

设 $(U_\alpha, \varphi_\beta)$ 是一个局部图 (local chart)，则 $\varphi_\alpha : U_\alpha \to \mathcal{Z}$ 称为编码映射 (encoding map)，参数域 $\mathcal{Z}$ 称为隐空间 (latent space) 或特征空间 (feature space)，其也是一个 Polish 空间，$\mathcal{Z}$ 的维度远低于 $\mathcal{X}$ 的维度。点 x $\in \Sigma$ 被称为样本，其参数 φ_α(x) 被称为 x 的特征 (feature)。此外，编码映射 $\varphi_\alpha : U_\alpha \to \mathcal{Z}$ 能够诱导出一在特征空间 $\mathcal{Z}$ 上定义的前推 (push-forward) 概率测度 $\varphi_\alpha^* \mu$。

φ_α 的逆映射 $\psi_\alpha := \varphi_\alpha^{-1} : \mathcal{Z} \to \Sigma$ 称为解码映射 (decoding map)，其给出了流形的局部参数表示。当然，编码映射并非一定可逆 (并且大部分情况下都是不可逆的)，若不可逆，则解码后的图片与编码前的图片可能不同，具体情况请见第五章。

第二节　Monge 问题与 Kantorovich 问题

最优传输理论来自于法国数学家蒙日 (Monge) 在 1781 年提出的 Monge 最优传输问题[19]：

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

$$\begin{aligned} &\min_T \int_X c(x, T(x)) d\mu(x) \\ &s.t. \quad T^* \mu = \nu \end{aligned} \tag{3.3}$$

Definition 6. 最优传输映射：*Monge* 问题的解 T 称为从 μ 到 ν 的最优传输映射。

Definition 7. 传输计划 *(transportation plan)*：一个 $X \times Y$ 上的联合测度 $\gamma(A \times B)$，其描述从 A 被分到 B 的份数。

然而直接求解 Monge 最优传输问题存在一定难度，我们在实际应用中通常求解下面介绍的 Kantorovich 问题。

Definition 8. *Kantorovich* 问题：$\mu \in \mathcal{P}(X)$，$\nu \in \mathcal{P}(Y)$，给定代价函数 $c(x, y)$。$\mathcal{A}_{opt}(\mu, \nu)$ 为所有满足 $\gamma(A \times Y) = \mu(A), \gamma(X \times B) = \nu(B)$ 的最优传输计划的全体 *(*即满足 $\pi_x^* \gamma = \mu_\theta, \pi_y^* \gamma = \nu$ 的概率测度 $\gamma \in \mathcal{P}(X \times Y)$ 的全体*)*，则最优传输的 *Kantorovich* 问题 *(*又称 *Kantorovich* 公式*)* 为：

$$\min_{\gamma \in \mathcal{A}_{opt}} \int_{X \times Y} c(x, y) d\gamma(x, y) \tag{3.4}$$

Definition 9. *Wasserstein* 距离：给定 *Kantorovich* 问题 *(3.4)*，使传输代价最小的解 $W_c(\mu, \nu)$ 即为 *Wasserstein* 距离，即

$$W_c(\mu, \nu) = \min_{\gamma \in \mathcal{A}_{opt}} \int_{X \times Y} c(x, y) d\gamma(x, y) \tag{3.5}$$

形象地说，如果将分布解释为在一个给定区域内堆积给定数量的泥土的两种不同方式，则 Kantorovich 问题是求解将一种泥土 (earth) 堆积方式变成另一种泥土堆积方式的最小成本；其中成本假定为移动的泥土量乘以移动的距离。因此 Kantorovich 问题的解 (Wasserstein 距离) 也被称为 Earth-Mover 距离[40]，简称 EM 距离。

我们不加证明地给出如下定理，具体证明请见 [4]：

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

$$\inf(\textit{Monge 问题}) = \min(\textit{Kantorovich 问题}) \tag{3.6}$$

由此可以知道 *Kantorovich* 问题可以看成 *Monge* 问题的一个松弛 *(relaxation)*[15]。

Kantorovich 问题的一些优势[15][16][17]：

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

我们在项目成果论文《最优传输理论与其几何基础》中已经证明了求解 Kantorovich 问题的办法是理论上可行的，即理论上一定存在 c-concave 函数 φ 使得 $supp(\gamma) \subset \partial^{c+}\varphi$。我们的算法就使用了离散状况下 Kantorovich 问题的求解，利用了 Kantorovich 问题的良好性质，尤其是它的对偶性。

第三节　Kantorovich 对偶问题

由 Kantorovich 问题的定义，我们可以看出 Kantorovich 问题 (尤其是离散情况) 实际上是一个线性规划问题，由线性规划问题的对偶理论[41]，我们可以研究它的对偶问题。

Definition 10. 令 $\mu \in \mathcal{P}(x), \nu \in \mathcal{P}(Y)$，则 *Kantorovich* 问题的对偶问题为[4]：

$$\begin{aligned} &\max_{\varphi \in L^1(\mu), \psi \in L^1(\nu)} \{\int_X \varphi(x) d\mu(x) + \int_Y \psi(y) d\nu(y)\} \\ &s.t. \quad \varphi(x) + \psi(y) \leq c(x,y) \end{aligned} \tag{3.7}$$

经过一些数学上的推导 (请见我们的项目成果论文《最优传输理论与其几何基础》

)，我们可以得到以下等式

$$\begin{aligned}\inf_{\gamma}\int c(x,y)d\gamma(x,y) &= \sup_{\varphi}\{\int \varphi(x)+\varphi^{c+}(y)d\gamma(x,y)\} \\ &= \sup_{\varphi}\{\int \varphi(x)d\mu(x)+\int \varphi^{c+}(y)d\nu(y)\}\end{aligned} \tag{3.8}$$

虽然略去了一些数学证明，为了提高可读性我们仍需要介绍几个概念，它们会在下面的数学说明中出现。

以下的 c 均代表代价函数 $c(x,y)$:

Definition 11. *c*-循环单调 *(c-cyclical monotone)*[15]：称 $\Gamma \subset X \times Y$ 是 *c*-循环单调的，如果满足条件：对于 $(x_i,y_i)\in\Gamma, 1\le i\le N$，下式对于所有置换 σ 成立：

$$\sum_{i=1}^{N} c\left(x_i,y_i\right) \le \sum_{i=1}^{N} c\left(x_i,y_{\sigma(i)}\right) \tag{3.9}$$

Definition 12. *c*-凹 *(c-concave)* 和 *c*-凸 *(c-convex)* 函数[15][16][17]：

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

Theorem 2. 对偶定理 *(Duality Theorem)*[4]：如果对偶问题的最大值可以被取到，则 (φ,ψ) 可以表示成 (φ,φ^{c+}) 形式，其中 φ 为 *c-concave* 函数。

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

第四章　最优传输理论的凸几何角度理解

由定理 (3.6) 我们知道 Kantorovich 问题可以看成 Monge 问题的一个松弛。当 $c(x,y)$ 是连续的，且满足 $\mu(x)=0,\forall x\in X$ 时，有 inf(Monge 问题) = min(Kantorovich 问题)。

然而最优传输问题并不一定在所有情况下都存在解。我们可以证明[4]，若 $c(x,y)=h(x-y)$，其中 h 是严格凸函数，则一定存在最优传输映射。而对于一般的代价函数，如 L^2 代价，由于它是严格凸函数，因此最优传输问题一定有解。

下面我们将会对 Kantorovich 方法以及 Brenier 方法进行说明。Kantorovich 方法从最优传输理论的角度对 WGAN 的原理进行了解释，而 Brenier 方法则是从凸几何的角度进行解释。我们给出在完全离散的情况下两种方法在 L^2 代价函数下的等价性，并将其进一步推广到半连续情况，从而我们可以在几何的观点下理解最优传输问题。最后再陈述最优传输映射的存在性以及其与 Kantorovich 势函数 $\varphi(x)$ 的等价性，为之后对 GAN 的数学解释以及我们基于 GAN 所提出的 OTVAE 做好理论准备。

第一节　Kantorovich 方法与 Brenier 方法

假设 μ 有在 X 上的紧支撑 Ω(Ω 是 X 上的凸域)，即

$$\Omega=supp(\mu)=\{x\in X|\mu(x)>0\} \tag{4.1}$$

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

$$\int_{\Omega}d\mu(x)=\sum_{i=1}^{k}v_i \tag{4.2}$$

4.1.1　Kantorovich 方法

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

$$\int_{Y}\varphi d\nu=\sum_{i=1}^{k}\varphi_j v_j \tag{4.3}$$

ϕ 的 c-变换由下式给出：

$$\varphi^c(x)=\min_{1\le j\le k}\{c(x,y_j)-\varphi_j\} \tag{4.4}$$

这包括一个 X 的胞腔剖分 (cell decomposition)：

$$X = \bigcup_{i=1}^{k} W_i(\varphi) \tag{4.5}$$

每个胞腔有

$$W_i(\varphi) = \{x \in X | c(x, y_i) - \phi_i \leq c(x, y_i) - \phi_j, \forall 1 \leq j \leq k\}. \tag{4.6}$$

根据 Wasserstein 距离的定义 (3.9) 和公式 (3.5)，我们可以如下定义能量：

$$E(\varphi) = \int_X \varphi^c d\mu + \int_Y \varphi d\nu \tag{4.7}$$

然后我们可以得到

$$E_D(\varphi) = \sum_{i=1}^{k} \varphi_i \left(\nu_i - w_i(\varphi)\right) + \sum_{j=1}^{k} \int_{W_j(\varphi)} c\left(x, y_j\right) d\mu \tag{4.8}$$

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

$$w_i(\varphi) = \mu\left(W_i(\varphi)\right) = \int_{W_i(\varphi)} d\mu(x) \tag{4.9}$$

则 μ 和 ν 间的 Wasserstein 距离有

$$W_c(\mu, \nu) = \max_{\varphi} E(\varphi) \tag{4.10}$$

4.1.2 Brenier 方法

Kantorovich 对偶方法适用于普通代价函数。而当成本函数是 L^2 范数下的距离 $c(x, y) = |x - y|^2$ 时，我们可以直接应用 Brenier 方法[24].

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

$$u_h(x) = \max_{i=1}^{k} \{\langle x, y_i \rangle + h_i\} \tag{4.11}$$

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

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

$$\Omega = \bigcup_{i=1}^{k} W_i(h) \tag{4.12}$$

其中每个胞腔 $W_i(h)$ 是 u_h 在 Ω 上一个面的投影。

$$W_i(h) = \{x \in X \mid \nabla u_h(x) - y_i\} \cap \Omega \tag{4.13}$$

$W_i(h)$ 的度量由下式给出：

$$w_i(h) = \int_{W_i(h)} d\mu \tag{4.14}$$

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

$$\nabla u_h : W_i(h) \to y_i, i - 1, 2, \cdots, k \tag{4.15}$$

将每个 $W_i(h)$ 映射到单个点 y_i。根据 Alexandrov 定理[26] 和 Gu-Luo-Yau 定理[3]，我们得出以下结论，具体证明请见 [3][4]：

Theorem 3. 设 Ω 是 $\mathbb{R}^n$ 中的紧凸域，$\{y_1, ..., y_k\}$ 是 $\mathbb{R}^n$ 中的一组不同点，μ 是 Ω 上的概率度量。对于任何 $\nu = \sum_{i=1}^{k} \nu_i \delta_{y_i}$，且 $\sum_{i=1}^{k} \nu_i = \mu(\Omega)$，存在 $h = (h_1, h_2, ..., h_k) \in \mathbb{R}^k$，在不考虑常数 $(c, c, ..., c)$ 的情况下，有 $w_i(h) = \nu_i$ 对于所有的 i 都成立。向量 h 是凹函数

$$E_B(h) = \sum_{i=1}^{k} h_i \nu_i - \int_0^h \sum_{i=1}^{k} w_i(\eta) d\eta_i \tag{4.16}$$

的最大点.

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

$$\int_{\Omega} |x - T(x)|^2 d\mu \tag{4.17}$$

因此 ∇u_h 有某种“最优性”，实际上，最优传输映射 $T = \nabla u_h$。这里我们就将 *Kantorovich* 方法与 *Brenier* 方法联系了起来。

第二节　最优传输映射的存在性与 Kantorovich 势函数的等价性

为了利用最优传输映射进行图片生成，我们必须保证其存在性。事实上，如果代价函数是凸函数 (很多常用的代价函数都满足该性质)，则最优传输映射一定存在。我们给出如下两个定理：

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

则 $\nabla\varphi(x_0) = \nabla_x c(x_0, y_0)$

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

$$\nabla\varphi(x_0) = \nabla_x c(x_0, y_0) \tag{4.18}$$

□

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

证明. 由定理 4，

$$\nabla\varphi(x_0) = \nabla_x c(x_0, y_0) = \nabla h(x_0 - y_0), \tag{4.19}$$

因为 h 是严格凸的，因此 ∇ 是可逆的.

$$x_0 - y_0 = (\nabla h)^{-1}(\nabla\varphi(x_0)), \tag{4.20}$$

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

定理 5 说明：当 h 是严格凸函数的时候，最优传输映射 T 一定存在，且其与 Kantorovich 势函数 $\varphi(x)$ 可以互相推导，因此最优传输映射和 Kantorovich 势函数是等价的。

第三节 L^2 代价下 Kantorovich 方法与 Brenier 方法的等价性

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

$$T(x) = x - \nabla\phi(x) = \nabla(\frac{x^2}{2} - \varphi(x)) = \nabla u(x) \tag{4.21}$$

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

$$u_h(x) = \frac{1}{2}|x|^2 - \varphi(x) \tag{4.22}$$

在半连续情况下，上式仍然成立，具体证明较为复杂，涉及到 Minkowski 问题[25] 和 Alexandrov 问题[26]，在此不展开赘述，具体证明可见 [3]，[4]。公式 (4.22) 说明在 L^2 代价下，Kantorovich 方法与 Brenier 方法求解的东西是可以互相转换的，因此我们称 Kantorovich 方法与 Brenier 方法有等价性。由于我们可以用最优传输理论解释 WGAN 中的 Wasserstein 距离，我们也可以用几何的方法来对 WGAN 进行一些解释。

第五章　生成模型的几何解释

在得到了最优传输映射 T 和 Kantorovich 势函数 $\varphi(x)$ 的等价性后，我们便有了进一步分析 GAN 所蕴含的数学原理的工具——微分几何。下面首先介绍一个在机器学习领域被较多人认可的流形假设，这是我们对 GAN 进行数学解释的起点。以此为前提，在我们上一章的结论的基础上，我们将会得到一个并不明显但符合直觉的结论：在 L^2 成本函数下，生成器和判别器的训练是相互促进的。这提示了我们应该在他们之间共享部分计算内容来提高效率，而不是进行纯“对抗”式的训练。最后，我们对自编码器算法也进行了数学方面的解释，为我们的模型 OTVAE 的可解释性提供了有力的支持。

第一节　机器学习中的流形假设

流形假设[2] 是机器学习中流形学习[42] 与非线性维度降低问题[43](如常见的聚类问题)的理论基础。其认为环境空间 $\mathcal{X}$ 指的是某一种自然界的数据 (如各种类型的图片) 所在的空间，其是极高维的空间。同样类型的数据 (如人脸图片) 在 $\mathcal{X}$ 中的一个流形 Σ 之上，其服从一个概率分布 μ。流形假设还认为：同样类型的数据的特征 (如人脸特征) 集中在特征空间 $\mathcal{Z}$(其维度比 $\mathcal{X}$ 低很多) 中的一个非线性低维流形附近，特征向量的分布为 ν。各种深度学习方法实际上是在从实际数据中学习流形结构并获得流形的参数表示。

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

第二节　GAN 的数学解释

生成器 G 可以看作一个从隐空间 $\mathcal{Z}$ 到样本空间 $\mathcal{X}$ 的一个映射 ($g_\theta : \mathcal{Z} \to \mathcal{X}$)，其中隐空间 $\mathcal{Z}$ 是样本空间 $\mathcal{X}$ 的特征空间。我们需要用深度网络学习参数 θ。

令 ζ 为隐空间 $\mathcal{Z}$ 上的一个简单分布 (例如高斯分布)，则生成器 G 使 ζ 变成 $\mu_\theta = g_\theta^* \zeta$ (其中 g_θ^* 为映射 g_θ 的导出映射)，μ_θ 就是样本空间 $\mathcal{X}$ 中一个由隐空间 $\mathcal{Z}$ 上的一个简单分布 ζ 导出的分布，这就与最优传输理论产生了联系。

判别器 D 利用某种距离计算生成分布 μ_θ 和训练集样本分布 ν 之间的误差，它由另一个深度神经网络通过参数 ξ 确定。如 WGAN 利用 Wasserstein 距离 $W_c(\mu_\theta, \nu)$ 来衡量生成分布和训练集分布的差距，由定义 9 和 13，Wasserstein 距离 $W_c(\mu_\theta, \nu)$ 等价于最优传输当中的 Kantorovich 势函数 φ_ξ。

总结来说，生成器 G 通过优化参数 θ，使得 μ_θ 逼近 ν；判别器 D 通过优化参数 ξ，使得 φ_ξ 逼近 Wasserstein 距离 $W_c(\mu_\theta, \nu)$。

用数学公式可以表示为以下的优化过程，这与公式 (2.1) 是等价的：

$$\min_{\theta}\max_{\xi}\mathbb{E}_{x\sim\zeta}\left(\varphi_{\xi}\left(g_{\theta}(x)\right)+\mathbb{E}_{y\sim\nu}\left(\varphi_{\xi}^{c}(y)\right)\right) \tag{5.1}$$

在最优传输的观点下[4][5]，生成对抗网络的生成器 G 中的映射 g_θ 等价于最优传输问题中的最优传输映射 T；生成对抗网络的判别器 D 中求解 Wasserstein 距离 $W_c(\mu_\theta,\nu)$，等价于求解最优传输问题中的 Kantorovich 势函数 φ_ξ。由上一章最后得出的结论，我们便得出最开始的结论：在凸的代价函数下，生成器和判别器的工作有等价性，因此生成器和判别器的训练是相互促进的。

第三节　自编码器的数学解释

自编码器 [6] 通常用于无监督学习，它被应用于压缩、去噪、预训练等[7][8]。依照流形假设，我们可以如下从几何角度解释自编码器[5]：自编码器学习数据的低维特征，将其表示为参数多面体流形，即从特征空间 (参数域) 到环境空间的分段线性 (Piecewise Linear，简称为 PL) 映射，PL 映射的像是一 $\mathcal{X}$ 中的高维多面体流形。然后自动编码器利用该多面体流形作为原分布流形在各种应用中的数据逼近。

5.3.1　自编码器的流形解释

自编码器是一种前馈的、非递归，输出层与输入层具有相同的节点数的神经网络。自编码器的能力是重建自身的输入。自编码器的隐藏层为一瓶颈层 (bottleneck)，其维度显著低于输入层和输出层，功能是降维。

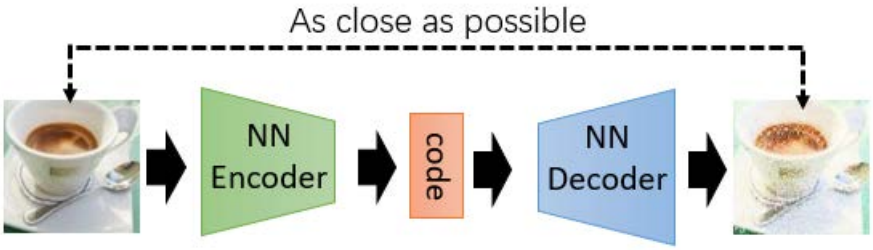

图 2: 自编码器的模型[27]

参照流形假设，我们可以有如下的记号：自编码器的输入空间是环境空间 $\mathcal{X}$，输出空间也是环境空间 $\mathcal{X}$。瓶颈层的输出空间是特征空间 $\mathcal{Z}$。自编码器通常由编码器 (encoder)φ 和解码器 (decoder)ψ 两部分组成。编码器获取一个样本 x $\in \mathcal{X}$ 并将其映射到 $z\in\mathcal{Z}, z=\varphi(x)$，图像 z 通常被称为 x 的潜在表示 (latent representation)。编码器 $\varphi:\mathcal{X}\to\mathcal{Z}$ 将 Σ 映射到它的潜在表示 $D=\varphi(\Sigma)$。之后，解码器 $\psi:\mathcal{Z}\to\mathcal{X}$ 将 z 映射到

与 x 维度相同的重构 $\tilde{\mathbf{x}}$，$\tilde{x} = \psi(z) = \psi \circ \varphi(x)$。

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

$$\varphi, \psi = \text{argmin}_{\varphi,\psi} \int_{\mathcal{X}} \mathcal{L}(x, \tilde{x}) d\mu(\mathrm{x}) \tag{5.2}$$

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

5.3.2 自编码器的实际应用

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

$$\min_{\theta} \mathcal{L}(\theta) = \min_{\theta} \frac{1}{k} \sum_{i=1}^{k} \left\| \mathrm{x}^{(i)} - \psi_\theta \circ \varphi_\theta \left(\mathrm{x}^{(i)}\right) \right\|^2 \tag{5.3}$$

其中编码器 φ_θ 和解码器 ψ_θ 都是 PL 映射。编码器 φ_θ 包含环境空间的单元分解 $\mathcal{D}(\varphi_\theta)$ [5]：

$$\mathcal{D}(\varphi_\theta) : \mathcal{X} = \bigcup_{\alpha} U_\theta^\alpha \tag{5.4}$$

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

5.3.3 用自编码器进行图片生成

假设 $\mathcal{X}$ 是所有 $n \times n$ 个图像的空间，其中每个点代表一个图像。我们可以定义一个概率测度 μ 表示图像表示要学习内容的分布。这里用人脸举例，人脸的形状是由有限数量的基因决定的，人脸照片是由人脸的几何结构、光线、摄像机参数等决定的。因此，假设所有的人脸照片都集中在一个有限维流形上是合理的，我们称之为人脸照片流形 Σ。通过使用大量真实的人脸照片，我们可以训练一个自编码器来学习人脸照片流形。学习过程产生解码映射 $\psi : \mathcal{Z} \to \tilde{\Sigma}$，即重构流形的参数表示。我们从一个简单分布 (如正态分布) 中随机生成一个向量 $z \in \mathcal{Z}$，再用正态分布到人脸特征分布的最优传输算法 T 把它投射到 $\mathcal{Z}$ 上，得到 x，则 x 应距离人脸特征分布流形较近，$\varphi(x) \in \tilde{\Sigma}$ 会以大概率给出一个人脸图像，而该图像很可能与训练集中的图像均不相同。因此这可以作为生成人脸照片的生成模型。

具体到我们的实验中，我们首先使用了手写阿拉伯数字集 [9] 进行训练。比起人脸数据集，阿拉伯数字的特征更少，相同数字之间的共性大于手写的随机性，同时随机因

素也较少，这样在训练上更简单，同时我们可以通过比较相同训练波数 (epoch) 下传统 GAN 和我们算法的生成结果质量，或通过比较生成相同质量的结果需要的波数，来体现出模型的优劣。

第六章　基于最优传输理论的生成模型

以上就是我们的 OTVAE 算法所需要的数学基础和主要设计思路，下面介绍 OTVAE 算法，首先介绍作为 OTVAE 的重要组成的变分自编码器的基本结构，再给出求取最优传输矩阵的算法，在这基础上给出 OTVAE 的算法，最后把它和公认有良好表现的 WGAN 进行比对，并给出比对结果。

第一节　变分自编码器 (VAE) 的结构

在我们的算法之前首先介绍一下变分自编码器 (VAE)[38]。VAE 是一种针对自编码器的改进算法。相比自编码器使用两组神经网络分别作为编码器以及解码器的结构而言，变分解码器生成特征向量的方法更为巧妙，同时拥有了一定的生成能力。我们这里主要利用 VAE 来提取特征向量以及学习生成器函数。

VAE 在原自编码器结构的基础上，改变了学习的内容，从学习特征向量变为对于每个样本单独学习一组正态分布。首先我们有一组数据样本 $X_{train} = \{x_1, x_2, ..., x_n\}$，来自 $\mathcal{X}$ 的图像的空间。我们从 X_{train} 中提取特征，按照之前的理论推导这里的特征向量满足分布 $\mu \in \mathcal{Z}$。给定一个图片样本 x_k，我们定义 $p(z|x_k)$ 是属于 x_k 的后验分布。假设其服从正态分布，我们用两个自编码器学习它的两个参数：均值向量 μ 与方差 σ^2。从这一后验分布 $p(z|x_k)$ 中进行采样，再训练一个生成器 $g(Z)$，把从分布 $p(z|x_k)$ 采样出来的一个 z_k 还原成 x_k。于是我们构建两个神经网络，分别学习 $\mu_k = f_1(x_k)$，$log\sigma_k^2 = f_2(x_k)$，按照上面的过程进行还原，再最小化损失函数。

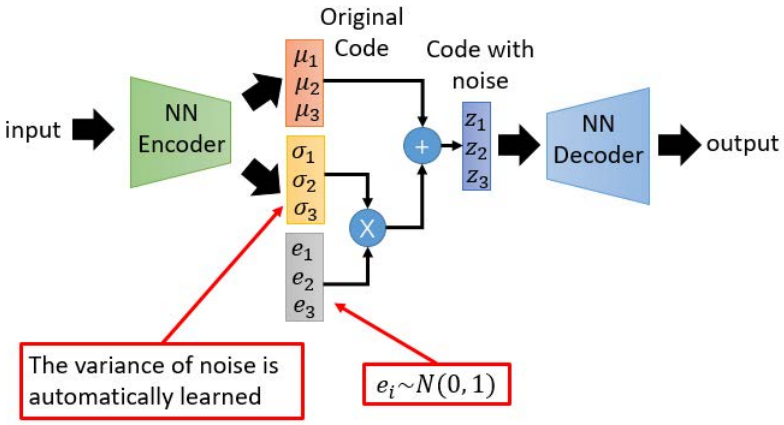

图 3: VAE 的模型[27]

下面是 VAE 的一个算法伪代码[38]：

Algorithm 3 VAE

Require: 训练集 $X_{train} = \{x_1, x_2, ..., x_n\}$, 对应的标签 $L = \{l_t\}_{t=1}^{T_x}$, 损失权重 $\lambda_1, \lambda_2, \lambda_3$

将 θ_0, ϕ_0 初始化

repeat

- 选取在这一小批量中的样本 x_t
- 解码器：$\mu_{z_t} \to f_\phi(x_t)$
- 抽样：$\mathbf{z}_t \leftarrow \boldsymbol{\mu}_{\mathbf{z}_t} + \boldsymbol{\epsilon} \odot \boldsymbol{\sigma}_{\mathbf{z}}, \boldsymbol{\epsilon} \sim \mathcal{N}(0, 1)$
- 生成器：$\mu_{x_t} \to f_\theta(z_t)$
- 计算重建损失:

$$\mathcal{L}_{rec} = -\log p_\theta\left(\mathbf{x}_t \mid \mathbf{z}_t\right) = -\log \mathcal{N}\left(\mathbf{x}_t; \boldsymbol{\mu}_{\mathbf{x}_t}, \boldsymbol{\sigma}_{\mathbf{x}}^2 \mathbf{I}\right)$$

- 计算正则损失:

$$\mathcal{L}_{reg} = \frac{1}{2}\left\|\mu_{\mathbf{z}_t}\right\| + \frac{1}{2}\left\|\sigma_{\mathrm{z}}\right\| - \frac{1}{2}\sum_{k=1}^{d}\left(1 + \log \sigma_{\mathbf{z}(k)}^2\right)$$

- 计算分类损失:

$$\mathcal{L}_{cls} = \operatorname{softmax}\left(\operatorname{loss}\left(\mathbf{z}_t, l_t\right)\right)$$

- 合并损失:

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}) = \lambda_1 \mathcal{L}_{rec}(\boldsymbol{\theta}, \boldsymbol{\phi}) + \lambda_2 \mathcal{L}_{reg}(\boldsymbol{\phi}) + \lambda_3 \mathcal{L}_{cls}(\boldsymbol{\phi})$$

- 梯度反向传播

until 达到最大迭代次数

我们通过 VAE 得到所需要的生成器以及特征向量空间。注意这里 VAE 本身也是一个生成模型，会生成与原特征向量有细微差别的特征向量。我们的算法使用最优传输映射从正态抽取随机向量再映射到特征空间，最后输入解码器。这样能让输入解码器的向量更贴近 μ 分布中的向量，使得生成的图像更有可能在该类图片的流形 Σ 上。

第二节　离散 Kantorovich 问题与 Sinkhorn 算法

6.2.1　离散 Kantorovich 问题

在我们的模型中，我们考虑的是离散形式的 Kantorovich 问题，也就是说我们要考虑的是从一个离散概率分布向量到另一个离散概率分布向量的最优传输问题。下面先

给出两个定义。

Definition 14. 耦合矩阵集合 *(coupling matrix set)*[18]：

$$U(r,c)=\{P\in\mathcal{R}_{+}^{n*m}|P1_m=r,P^T1_n=c\} \tag{6.1}$$

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

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

$$d_M(r,c)=\min_{P\in U(r,c)}P\odot M==\min_{P\in U(r,c)}\sum P_{ij}M_{ij} \tag{6.2}$$

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

我们要做的就是求解最优传输矩阵 P，其与最优传输映射 T 在 r 与 c 维数相等时是等价的。由定义 15 可见，最优传输映射和 Wasserstein 距离是等价的，我们求解最优传输映射，就是求解 Wasserstein 距离，这又一次印证了 GAN 中生成器和判别器的促进性，这也是我们抛弃生成器和判别器而进行直接生成的原因。

我们用一个例子通俗地解释一下离散的最优传输问题[20]：

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

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

每个同事对每种小吃的喜好程度矩阵为 M，M 给定。

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

6.2.2 Sinkhorn 算法

Sinkhorn 算法有无正则化和有正则化两种不同形式，其中正则化[22] 会让分配更加趋向于平均分配[18]。我们下面采用无正则化的 Sinkhorn 算法，使用的是 Python 中 POT 包[21] 的 ot.emd() 函数，其算法为运筹学中非常基础的单纯形法，在此并不赘述，具体算法请见 [23]。

我们知道离散 Kantorovich 问题一定有解。Sinkhorn 算法证明[13]：无论是否有正则化问题，都可以通过迭代的方法求解 P 的近似值。这是因为在满足原问题的最优传输条件以及保测度条件后，可以把原问题转化成一组属于矩阵放缩的数学问题的等式，每一步分别满足一个等式，最终迭代一定会收敛。

下面介绍求解离散 Kantorovich 问题的 Sinkorn 算法：

设特征空间 $\mathcal{Z}$ 为 $dim_{\mathcal{Z}}$ 维空间 (在我们的算法中 $dim_{\mathcal{Z}} = 32$)，n 为训练集中图片数量。所有的训练集特征向量记为 $x_i(i = 1, ..., n)$，这些向量的经验分布为 $\mu \in P(Z)$，它们两两不同。从 $dim_{\mathcal{Z}}$ 维标准正态分布中抽取 n 个随机向量 $y_j(j = 1, ..., n)$，这些向量的经验分布为 $\zeta \in P(Z)$(这里 ζ 是对正态分布的估计)，它们两两不同。

由于 x_i 两两不同，所以 μ 作为经验分布，一定有 $P(x = x_i) = 1/n, i = 1, ..., n$，同理 ζ 为 $P(y = y_j) = 1/n, j = 1, ..., n$，这样二者的边缘分布向量 r 和 c 都等于 $(1/n, ..., 1/n)$(n 维)。下面以 ζ 和 μ 的距离 (我们采用欧式距离) 为代价函数，解决无正则化的离散 Kantorovich 问题。

Algorithm 4 Sinkhorn 算法求 ζ 到 μ 的最优传输矩阵 P

Require: $x_i(i = 1, ..., n), y_j(j = 1, ..., n)$

Ensure: 最优传输矩阵 P

- 生成一个 n 维向量 $c = \frac{1}{n}, \frac{1}{n}, ..., \frac{1}{n}$
- 生成一个 n 维向量 $r = \frac{1}{n}, \frac{1}{n}, ..., \frac{1}{n}$
- 使用 L^2 范数计算 ζ 和 μ 之间的距离矩阵 M，其便是代价矩阵
- 对 M 进行归一化
- 用单纯形法 [23] 求得最优传输矩阵 P

我们这里使用了 POT 包[21] 的 ot.emd() 函数来进行单纯形法的求解，最终求出 c 到 r 的最优传输矩阵 P。由于 r 和 c 维数相等，最优传输映射是一个双射[14]，则最优传输矩阵是一个每行每列都有且只有一个 $\frac{1}{n}$ 的 n 维方阵。最优传输矩阵中所有非零的位置 (如 a_{ij}) 代表 T(第 j 个 ζ 中的向量)= 第 i 个 μ 中的向量，注意此时 T 的定义域是 ζ 里面所有的向量，是一个有限的集合，也就是说在应用这个最优传输矩阵时我们需要进行延拓。

第三节　新的生成模型 (OTVAE)

我们希望把 Sinkhorn 算法与 VAE 模型结合在一起，实现生成图片的功能。

我们继续使用上面的记号，此外，特征空间 $\mathcal{X}$ 为 $dim_{\mathcal{X}}$ 维空间 ($dim_{\mathcal{X}} \gg dim_{\mathcal{Z}}$)。$X_{train}$ 为训练集所有图片的集合，该训练集中的图片是同一类图片 (如手写数字)。再从 $dim_{\mathcal{Z}}$ 维正态分布中随机抽取 m(m 为想要生成的图片数量) 个向量，其集合记为 $\tilde{X}$。我们希望生成 m 个与原图片类型相同的图片，集合为 X_{gen}。

下面是 OTVAE 的伪代码。注意到我们这里使用了 Top-k 算法[44]，它是一种利用大根堆排序找到数组里的最小的 k 个数的算法。在我们的生成过程中，$dim_{\mathcal{X}} = 28 * 28 = 784, dim_{\mathcal{Z}} = 32, k = 2$。

Algorithm 5 OTVAE

Require: $X_{train}, \tilde{X}$

Ensure: 生成的图片集合 X_{gen}

- 用 X_{train} 按照算法 3 训练一个 VAE，其输入层、输出层维数 $dim_{\mathcal{X}}$, 瓶颈层维数为 $dim_{\mathcal{Z}}$，输入层到瓶颈层的映射 (编码映射)f，瓶颈层到输出层的映射 (解码映射)g。
- $f(X_{train})$ 得到 X_{train} 所有的特征向量，记为 $x_i(i = 1□...n)$，这些向量的经验分布为 μ。
- 从 $dim_{\mathcal{Z}}$ 维标准正态分布中抽取 n 个随机向量 $y_j(j = 1, ..., n)$，这些向量的经验分布为 ζ。
- 使用算法 4 求得 ζ 到 μ 的离散最优传输映射 T

for $\tilde{x}_i \in \tilde{X}$ **do**

- 用 Top-k 算法找出 ζ 中离 $\tilde{x}_i$ 的欧氏距离最近的 k 个向量 k_{min}
- 得到 T 作用在 k_{min} 上的向量组 $desired - vecs$
- 用 $dim_{\mathcal{Z}}$ 维 Dirichlet 分布[45] 生成一个元素总和为 1 的 k 维向量 $coef$
- 得到 $T(\tilde{x})$ 为 $desired - vecs$ 的线性组合，其系数为 $coef$ 的各分量。

end for

- 上面得到 m 个 $dim_{\mathcal{Z}}$ 维向量，记为 $z_k(k = 1, \ldots, m)$
- 用 $g(z_k)$ 得到由 z_k 生成的图片

中间的循环是为了把之前得到的最优传输映射 T 的定义域从离散的 ζ 里面的向量延拓到整个连续的空间上，令其变成一个连续的 $\mathcal{Z} \to \mathcal{Z}$ 映射，并保证输入任一正态分布随机向量，输出向量 μ 中抽出来的向量距离较近。

第四节　OTVAE 与 WGAN 的性能比较

6.4.1　性能度量与评价方法

不像传统的计算机视觉领域，生成模型没有严谨的性能度量。评价一个生成模型很大程度上靠人眼来认定其生成结果是否“真实”。这样的度量直观且简单，但太过主观，仅能作为一种参考。基于人眼判定，我们提出了“优秀率”的概念，以其作为一种仅供参考的性能度量。假定我们每训练一个 epoch，会输出 N 个数字，人眼认定较为“真实”(称为“优秀”) 的图片数量为 K，则记优秀率为优秀图片占图片总数的比例，即 $p = \frac{K}{N}$。

根据优秀率，我们便可以比较在同样优秀率的情况下，需要训练的 epoch 数量；也可以比较在训练的 epoch 数量相同的情况下，两个模型优秀率之间的差距。

6.4.2 实验结果

我们对 WGAN 进行了 200 个 epoch 的训练，采用的参数如下表所示：

参数	含义	值	参数	含义	值
n_epochs	训练的 epoch 上限	200	channels	通道数	1
batch_size	每次训练的批大小	100	n_critic	经过多少次 G 的训练更新一次 D	5
lr	学习率	0.00005	clip_value	权重裁剪的值	0.01
latent_dim	隐空间维度	100	sample_interval	取样之间的区间长度	100
img_size	图片的尺寸大小	28			

表 1: WGAN 的训练参数

对 OTVAE 进行了 100 个 epoch 的训练，采用的参数如下表所示：

参数	含义	值
pth_path	VAE 模型训练权重	”，即重新训练 VAE
k	随机生成所需要的邻居向量个数	2
epoches	每张图片生成数字的数量	10000
dim_z	隐空间维度	32

表 2: OTVAE 的训练参数

在最终状态下，二者生成的图片样本如下图所示：

图 4: WGAN 训练 200 个 epoch

图 5: OTVAE 训练 100 个 epoch

可以看到，WGAN 训练 200 个 epoch 时，优秀率只有 35%-40% 左右；而 OTVAE 训

练 100 个 epoch 就能达到 80% 以上的优秀率，事实上，OTVAE 训练 3-4 个 epoch，就能达到 40% 左右的优秀率了。

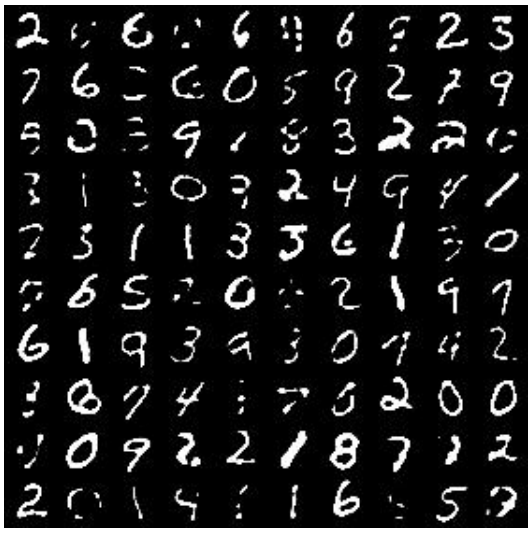

图 6: OTVAE 训练 3 个 epoch

综上所述，我们的模型无论是从生成质量还是生成速度上，都优于 WGAN。此外，由于我们的模型不基于 GAN，因此并不会产生模式崩溃的问题。这也是我们的模型优于 WGAN 很重要的一点。

我们将使用的 WGAN 和 OTVAE 的 Python 代码放于我们的 Github 仓库上，请访问 https://github.com/jimcui3/OTVAE。读者可在学习与研究的目的下自行取用并测试，若要进行商用，请联系作者。

第七章　总结与展望

在这篇论文中，我们先回顾了生成对抗网络 (GAN) 的模型、原理与优劣势，并介绍了其改进模型 WGAN。之后我们通过几何方法和最优传输理论对 GAN 及其他生成模型进行了数学上的解释。为此我们介绍了几何中的流形的知识与 Brenier 方法，最优传输理论中的 Monge 问题、Kantorovich 问题与 Kantorovich 方法。我们通过证明凸代价函数下最优传输映射与 Kantorovich 势函数的等价性与 L^2 代价下 Kantorovich 方法与 Brenier 方法的等价性，说明了 GAN 的生成器和判别器并非完全对立的，甚至是可以互相促进的。为了利用其促进关系，我们提出了基于最优传输映射和 VAE 模型的 OTVAE 模型，其也是一个生成模型，我们利用实验证明了其比传统的 WGAN 在速度和生成质量上更加优秀。

我们在建立 OTVAE 模型的过程中遇到了一些有趣的问题，由于时间问题，我们没有办法对其一一探究。例如：能否用更深的微分几何知识、最优传输知识甚至其他方向的数学知识解释、简化生成式模型，能否利用有正则化的 Sinkhorn 算法求解最优传输映射，能否利用高维插值等算法进行最优传输映射 T 的延拓，有没有进一步降低算法复杂度的方法，如果将 VAE 换为其他的模型会如何，有没有比“优秀率”评价生成式模型更严谨的性能度量，等等。针对这些问题，我们和其他的研究者可以在未来进行相关的研究。模型的可解释性是一个新兴而又难度较高的方向，我们做的只是学习众多前沿学者的成果并进行简单的尝试。若想实现进一步的重大突破，我们还需要学习更深的数学、计算机的知识，同时离不开其他研究者对此方向的继续研究。

参考文献

[1] B. Riemann (1867).

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

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

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

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

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

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

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

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

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

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

[12] Radford, A., Metz, L., and Chintala, S., Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, arXiv e-prints, 2015.

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

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

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

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

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

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

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

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

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

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

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

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

[25] Klain, Daniel A. (2004), The Minkowski problem for polytopes, Advances in Mathematics, 185 (2): 270–288, doi:10.1016/j.aim.2003.07.001, MR 2060470

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

[27] Li, H.. Website: https://speech.ee.ntu.edu.tw/ hylee/mlds/2018-spring.html, 2018

[28] Kullback, S. (1959), Information Theory and Statistics, John Wiley and Sons. Republished by Dover Publications in 1968; reprinted in 1978: ISBN 0-8446-5625-9.

[29] Ôsterreicher, F.; I. Vajda (2003). A new class of metric divergences on probability spaces and its statistical applications. Ann. Inst. Statist. Math. 55 (3): 639–653. doi:10.1007/BF02517812.

[30] Hastie, Trevor. Tibshirani, Robert. Friedman, Jerome. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, NY, 2009.

[31] Brock, A., Donahue, J., and Simonyan, K., Large Scale GAN Training for High Fidelity Natural Image Synthesis, arXiv e-prints, 2018.

[32] Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., and Efros, A. A., Context Encoders: Feature Learning by Inpainting, arXiv e-prints, 2016.

[33] Lei, N., Guo, Y., An, D., Qi, X., Luo, Z., Yau, S.-T., and Xianfeng Gu, D., (2019). Mode Collapse and Regularity of Optimal Transportation Maps.

[34] Arjovsky, M. and Bottou, L., Towards Principled Methods for Training Generative Adversarial Networks, arXiv e-prints, 2017.

[35] Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y., Spectral Normalization for Generative Adversarial Networks, arXiv e-prints, 2018.

[36] Keras, Website: https://keras.io/api/optimizers/rmsprop/

[37] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A., Improved Training of Wasserstein GANs, arXiv e-prints, 2017.

[38] Kingma, D. P. and Welling, M., Auto-Encoding Variational Bayes, arXiv e-prints, 2013.

[39] Gur, S., Benaim, S., and Wolf, L., Hierarchical Patch VAE-GAN: Generating Diverse Videos from a Single Sample, arXiv e-prints, 2020.

[40] Y. Rubner, C. Tomasi, and L. J. Guibas. A metric for distributions with applications to image databases. In IEEE International Conference on Computer Vision, pages 59-66, January 1998.

[41] Boyd, Stephen P.; Vandenberghe, Lieven (2004). Convex Optimization (pdf). Cambridge University Press. p. 216. ISBN 978-0-521-83378-3. Retrieved October 15, 2011.

[42] Belkin, Mikhail (August 2003). Problems of Learning on Manifolds (PhD Thesis). Department of Mathematics, The University of Chicago.

[43] John A. Lee, Michel Verleysen, Nonlinear Dimensionality Reduction, Springer, 2007.

[44] Das G. (2009) Top-k Algorithms and Applications. In: Zhou X., Yokota H., Deng K., Liu Q. (eds) Database Systems for Advanced Applications. DASFAA 2009. Lecture Notes in Computer Science, vol 5463. Springer, Berlin, Heidelberg.

[45] S. Kotz; N. Balakrishnan; N. L. Johnson (2000). Continuous Multivariate Distributions. Volume 1: Models and Applications. New York: Wiley. ISBN 978-0-471-18387-7. (Chapter 49: Dirichlet and In-

verted Dirichlet Distributions)