Advances in Applied Mathematics 应用数学进展, 2025, 14(5), 601-613
Published Online May 2025 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2025.145286

Hans汉斯

基于Wasserstein距离作为GAN的优化目标提高其训练稳定性的理论研究

张惠玲

上海理工大学理学院，上海

收稿日期：2025年4月28日；录用日期：2025年5月21日；发布日期：2025年5月30日

摘　要

生成对抗网络(Generative adversarial Nets，以下简称GANs)因其在图像生成等领域的成功应用而备受关注。然而，其训练的不稳定性一直是一个难以解决的问题，训练过程常常受到模式崩溃、梯度消失和优化不稳定的困扰。一般提高GANs训练稳定性的方法有替代损失函数、梯度惩罚、谱归一化、批量归一化和架构改进等方法。但是这些研究大多缺乏理论基础，未给出相对完善的理论证明，本论文的目标是深入理解基于Wasserstein距离训练GANs的不稳定性，提供较为完整的理论证明。并探讨了进一步改进WGAN训练稳定性的策略，如梯度惩罚(WGAN-GP)，以提高WGAN训练的稳定性和泛化能力。本文的主要研究内容如下：第一部分：分析了WGAN通过最小化Wasserstein距离(简称W距离)代替传统的Jensen-divergence (简称JS散度)，避免了梯度消失问题。其关键优势在于采用了1-Lipschitz连续的判别器，确保了在训练过程中生成器能够从判别器获得有效梯度。其次，证明了W距离相较于其他距离或者散度对于概率分布序列具有良好的连续性和收敛性。第二部分：通过引入W距离替代原来两个分布之间的JS散度，从理论上改善了GANs训练的稳定性。然而，WGAN的实现仍面临挑战，如权重裁剪导致的容量利用不足和梯度消失问题。为此，基于W距离，Gulrajani等人提出了梯度惩罚(WGAN-GP)来满足Lipschitz约束，以进一步提高训练稳定性。但是大多文献直接给出梯度惩罚常数为1，并未给出具体证明，在本文中给出了证明。

关键词

GANs训练稳定性，Wasserstein距离，1-Lipschitz连续，权重裁剪，梯度惩罚

A Theoretical Study of Improving the Training Stability of GAN Based on Wasserstein Distance as Optimization Objective

Huiling Zhang

文章引用: 张惠玲. 基于 Wasserstein 距离作为 GAN 的优化目标提高其训练稳定性的理论研究[J]. 应用数学进展, 2025, 14(5): 601-613. DOI: 10.12677/aam.2025.145286

张惠玲

College of Science, University of Shanghai for Science and Technology, Shanghai

Received: Apr. 28th, 2025; accepted: May 21st, 2025; published: May 30th, 2025

Abstract

Generative adversarial networks (GANs) have attracted much attention due to their successful applications in fields such as image generation. However, the instability of their training has always been a difficult problem to solve, and the training process is often plagued by mode collapse, gradient vanishing, and optimization instability. General methods to improve the stability of GANs training include alternative loss functions, gradient penalties, spectral normalization, batch normalization, and architectural improvements. However, most of these studies lack a theoretical basis and do not provide a relatively complete theoretical proof. The goal of this paper is to deeply understand the instability of GANs training based on Wasserstein distance and provide a relatively complete theoretical proof. It also explores strategies to further improve the stability of WGAN training, such as gradient penalty (WGAN-GP), to improve the stability and generalization ability of WGAN training. The main research contents of this paper are as follows: Part I: WGAN is analyzed to avoid the gradient vanishing problem by minimizing the Wasserstein distance (W distance for short) instead of the traditional Jensen-divergence (JS divergence for short). Its key advantage is the use of 1-Lipschitz continuous discriminator, which ensures that the generator can obtain effective gradients from the discriminator during training. Secondly, it is proved that W distance has good continuity and convergence for probability distribution sequences compared with other distances or divergences. Part II: By introducing W distance to replace the JS divergence between the original two distributions, the stability of GANs training is theoretically improved. However, the implementation of WGAN still faces challenges, such as insufficient capacity utilization and gradient vanishing problems caused by weight clipping. To this end, based on W distance, Gulrajani *et al*. proposed a gradient penalty (WGAN-GP) to meet the Lipschitz constraint to further improve training stability. However, most literature directly gives the gradient penalty constant as 1, without giving a specific proof, which is given in this article.

Keywords

GANs Training Stability, Wasserstein Distance, 1-Lipschitz Continuity, Weight Clipping, Gradient Punishment

1. 研究背景

生成对抗网络(GANs) [1]自 2014 年提出以来，在图像生成[2] [3]、数据增强、风格迁移、超分辨率重建等领域取得了显著成果。然而，尽管 GANs 在生成高质量数据方面表现优异，但其训练过程却极不稳定，面临诸多挑战，包括模式崩溃、梯度消失或爆炸、训练震荡以及难以收敛等问题[4]-[6]。这些问题严重影响了 GANs 在实际应用中的可靠性和推广性。

为了解决训练稳定性问题，研究者们提出了多种改进方法，如改进损失函数、引入正则化技术(梯度惩罚[7]、谱归一化 SN-GAN [8])、调整网络结构(自注意力机制 Self-Attention [9] GAN、StyleGAN [9])等。然而，GANs 训练稳定性的理论分析和实践优化仍然是一个开放性问题，值得进一步深入研究。

2. 标准 GANs 的原理及结构

GANs 是一类强大的生成模型，通过两个神经网络(生成器和鉴别器)之间的博弈过程进行训练，训练过程被建模为一个 Min-Max 问题，生成器(Generator)捕获数据分布、生成逼真的数据，以欺骗鉴别器，鉴别器(Discriminator)估计样本是来自真实分布还是生成分布的概率，通过彼此博弈，直到鉴别器输出概率稳定为 0.5，最终达到纳什平衡。

生成器(G)的核心是先从一个简单的先验 $z \sim p(z)$ 中采样(例如均匀分布或者高斯分布)，然后映射到样本空间 $g_\theta(z)$，有时最后会添加噪音。g_θ 是一个由 θ 参数化的神经网络。

鉴别器(D)接收生成器生成的样本或者真实数据样本，并且区分两者，输出一个概率值。样本来自真实数据输出 1，来自生成数据输出 0。

GANs 整体训练目标：

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim\mathbb{P}_r}\left[\log D(x)\right] + \mathbb{E}_{z\sim\mathbb{P}_g}\left[\log\left(1-D(z)\right)\right] \tag{2.1}$$

经典的 GANs 训练过程可以分为两步，首先固定生成器，训练鉴别器：

使

$$L(D,g_\theta) = \mathbb{E}_{x\sim\mathbb{P}_r}\left[\log D(x)\right] + \mathbb{E}_{z\sim\mathbb{P}_g}\left[\log\left(1-D(z)\right)\right] \tag{2.2}$$

达到最大。

关于 D 对 $L(D,g_\theta)$ 进行求导，在理论上[1]鉴别器可以收敛到最优状态为：

$$D^*(x) = \frac{P_r(x)}{P_r(x)+P_g(x)} \tag{2.3}$$

其次，当鉴别器达到最优时，生成器的优化目标则为：

$$\begin{aligned}
L(D^*,g_\theta) &= \int_{x\sim\mathbb{P}_r} P_r(x)\log D^*(x)\mathrm{d}x + \int_{z\sim\mathbb{P}_g} P_g(x)\log\left(1-D^*\left(g_\theta(z)\right)\right)\mathrm{d}z \\
&= \int_x \left[P_r(x)\log D^*(x) + P_g(x)\log\left(1-D^*(x)\right)\right]\mathrm{d}x \\
&= \int_x \left[\left(\log 2-\log 2\right)P_r(x) + P_r(x)\log\frac{P_r(x)}{P_r(x)+P_g(x)}\right. \\
&\quad \left. +\left(\log 2-\log 2\right)P_g(x) + P_g(x)\log\frac{P_g(x)}{P_r(x)+P_g(x)}\right]\mathrm{d}x \\
&= -\log 2\int_x \left(P_r(x)+P_g(x)\right)\mathrm{d}x + \int_x \left[P_r(x)\left(\log 2+\log\frac{P_r(x)}{P_r(x)+P_g(x)}\right)\right. \\
&\quad \left. + P_g(x)\left(\log 2+\log\frac{P_g(x)}{P_r(x)+P_g(x)}\right)\right]\mathrm{d}x \\
&= -2\log 2 + \int_x \left[P_r(x)\log\frac{2P_r(x)}{P_r(x)+P_g(x)} + P_g(x)\log\frac{2P_g(x)}{P_r(x)+P_g(x)}\right]\mathrm{d}x \\
&= -2\log 2 + KL\left(\mathbb{P}_r(x)\,\|\,\frac{\mathbb{P}_r(x)+\mathbb{P}_g(x)}{2}\right) + KL\left(\mathbb{P}_g(x)\,\|\,\frac{\mathbb{P}_r(x)+\mathbb{P}_g(x)}{2}\right) \\
&= -2\log 2 + 2JSD\left(\mathbb{P}_r(x)\,\|\,\mathbb{P}_g(x)\right)
\end{aligned} \tag{2.4}$$

通过以上分析可得，当鉴别器达到最优时，整个 GANs 的目标(2.1)就变为了最小化真实分布和生成

分布的 JS 散度。因此，理论上，期望首先将鉴别器训练得尽可能接近最优值，然后关于 θ 进行梯度下降，交替进行这两步。然而，在实践中，随着鉴别器训练得越来越好，生成器更新会变得越来越糟。这个问题主要是由损失函数饱和引起的。

3. WGAN 稳定性的理论证明

本章主要证明 WGAN 通过最小化 Wassertein 距离替代 JS 散度作为 GANs 训练的目标函数，避免了梯度消失的问题。WGAN 的关键优势在于它采用了 1-Lipschitz 连续的判别器(critic)，这确保了在训练过程中，生成器始终能够从判别器那里获得有效的梯度。这一改进使得 GANs 的训练变得更加稳定，并且由于 WGAN 的目标函数使得生成器不会专注于某些模式，而是更均匀地逼近真实分布，因此显著减少模式崩溃问题。

3.1. 问题描述

为了解决 JS 散度在训练中带来的问题，Arjovsky 等人[4] [5]引入了 Wassertein-1 距离作为衡量两个分布之间差异的指标。并采用 Kantorovich-Rubinstein 对偶性将其转化为一个可解的优化问题，其中对偶形式要求判别器(critic)是 1-Lipschitz 函数，这相当于在函数判别器上加了一个平滑约束，使其不会变化太快。

3.2. Wasserstein 距离的理论优越性

对比多种概率分布之间的距离衡量指标，包括 TV 距离、KL 散度、JS 散度和 W 距离。W 距离比其他距离或者散度具有更好的理论性质，提供有意义的梯度，可以正确衡量两个分布之间的距离。衡量分布之间的距离或者散度的不同指标最根本的区别在于它们对概率分布序列收敛的影响。

下面将具体说明 W 距离具有的优良性质：

首先通过一个示例说明概率分布序列如何在 W 距离下收敛，而在其他距离和散度下却不收敛。

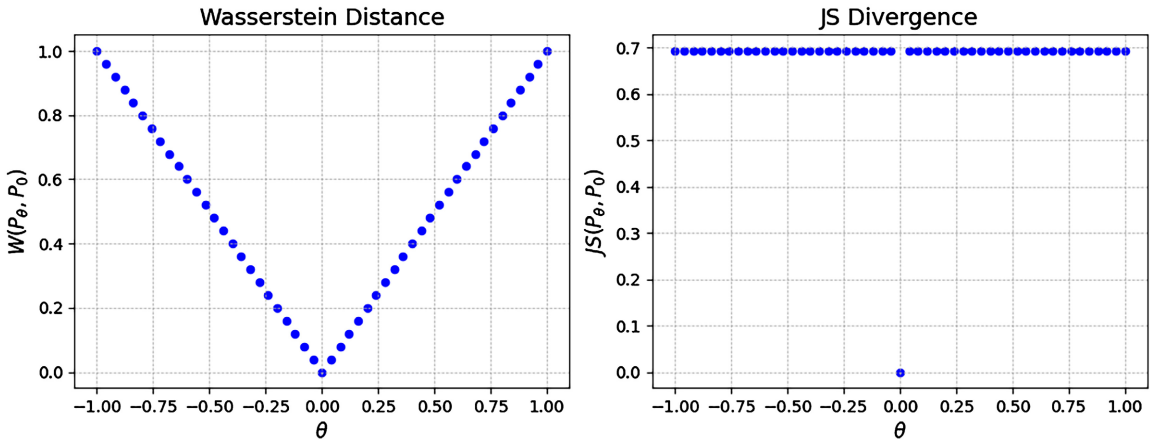

Figure 1. Function graphs of W distance and JS divergence with respect to θ
图 1. W 距离和 JS 散度关于 θ 的函数图像

例 3.1 设 $\mathbb{Z} \sim U[0,1]$ 是单位区间上的均匀分布。$\mathbb{P}_0$ 是二维空间 $(0,\mathbb{Z}) \in \mathbb{R}^2$ 上的分布(x 轴是 0，y 轴是随机变量 $\mathbb{Z}$)，在通过原点的垂直直线上均匀分布。令 $g_\theta(z) = (\theta, z)$，其中 θ 是单个实参数。在这种情况下，有：

- $W(\mathbb{P}_0, \mathbb{P}_\theta) = |\theta|$；

- $JS\left(\mathbb{P}_0,\mathbb{P}_\theta\right)=\begin{cases}\log 2, \theta\neq 0\\ 0, \theta=0\end{cases}$；
- $KL\left(\mathbb{P}_0\,\|\,\mathbb{P}_\theta\right)=KL\left(\mathbb{P}_\theta\,\|\,\mathbb{P}_0\right)=\begin{cases}+\infty, \theta\neq 0\\ 0, \theta=0\end{cases}$；
- $\delta\left(\mathbb{P}_0,\mathbb{P}_\theta\right)=\begin{cases}1, \theta\neq 0\\ 0, \theta=0\end{cases}$。

当$\theta_t\to 0$时，分布序列$\left(\mathbb{P}_{\theta_t}\right)_{t\in N}$在 W 距离下收敛，但在 JS 散度、KL 散度、逆 KL 散度和 TV 距离下均不收敛。如图 1 所示。

由以上示例可以看出，可以通过对 W 距离进行梯度下降来学习低维流形上的概率分布。而对于其他距离和散度，这是无法实现的，因为由此产生的损失函数是不连续的。

那么接下来我们考虑在简单的假设下，W 距离对于分布$\mathbb{P}_g$和$\mathbb{P}_r$是否也是θ上的连续损失函数：

定理 3.1 令$\mathbb{P}_r$为χ上的固定分布。令z为另一个空间$\mathcal{Z}$上的随机变量(例如高斯变量)。令$\mathbb{P}_g$表示$g_\theta(z)$的分布，其中$g:(z,\theta)\in\mathcal{Z}\times\mathbb{R}^d\mapsto g_\theta(z)\in\mathcal{X}$。则，

1) 如果g关于θ连续，则$W\left(\mathbb{P}_r,\mathbb{P}_g\right)$也关于$\theta$连续。

2) 如果g局部服从 Lipschitz 函数且满足上述假设 1)的规律性，则$W\left(\mathbb{P}_r,\mathbb{P}_g\right)$在各处连续，且几乎在各处可微。

3) 对于 Jensen-Shannon 散度$JS\left(\mathbb{P}_r,\mathbb{P}_g\right)$和所有 KL 散度，陈述 1)~2)都是错误的。

证明 以下将逐条证明。

1) 令θ和θ'为$\mathbb{R}^d$中的两个参数向量。首先可以约束$W\left(\mathbb{P}_{g_\theta},\mathbb{P}_{g_{\theta'}}\right)$，由此再证明定理。证明的主要元素是耦合$\gamma$的使用，即联合$\left(g_\theta(z),g_{\theta'}(z)\right)$的分布，显然有$\gamma\in\prod\left(\mathbb{P}_{g_\theta},\mathbb{P}_{g_{\theta'}}\right)$。

根据 W 距离的定义，有

$$\begin{aligned}W\left(\mathbb{P}_{g_\theta},\mathbb{P}_{g_{\theta'}}\right)&\le\int_{\chi\times\chi}\|x-y\|\mathrm{d}\gamma\\&=\mathbb{E}_{(x,y)\sim\gamma}\left[\|x-y\|\right]\\&=\mathbb{E}_z\left[\left\|g_\theta(z)-g_{\theta'}(z)\right\|\right]\end{aligned}\tag{3.1}$$

如果g关于θ连续，则$g_\theta(z)_{\theta\to\theta'}\to g_{\theta'}(z)$，故作为$z$的函数逐点有$\left\|g_\theta(z)-g_{\theta'}(z)\right\|\to 0$。由于$\chi$是紧集，那么其中任意两个元素的距离由固定常数$M$均匀地限制，因此对于所有的$\theta$和$z$，有$\left\|g_\theta(z)-g_{\theta'}(z)\right\|\le M$。通过有界收敛定理，有

$$W\left(\mathbb{P}_{g_\theta},\mathbb{P}_{g_{\theta'}}\right)\le\mathbb{E}_z\left[\left\|g_\theta(z)-g_{\theta'}(z)\right\|\right]\to_{\theta\to\theta'}0.\tag{3.2}$$

最后，可以得出：

$$\left|W\left(\mathbb{P}_r,\mathbb{P}_{g_\theta}\right)-W\left(\mathbb{P}_r,\mathbb{P}_{g_{\theta'}}\right)\right|\le W\left(\mathbb{P}_{g_\theta},\mathbb{P}_{g_{\theta'}}\right)\to_{\theta\to\theta'}0\tag{3.3}$$

即证明了$W\left(\mathbb{P}_r,\mathbb{P}_{g_\theta}\right)$关于$\theta$的连续性。

2) 令g服从局部 Lipschitz，那么对于给定的(θ,z)对，存在一个常数$L(\theta,z)$和一个开集U，使得$(\theta,z)\in U$，这样对于任意(θ',z')，都有

$$\left\|g_\theta(z)-g_{\theta'}(z')\right\|\le L(\theta,z)\left(\|\theta-\theta'\|+\|z-z'\|\right)\tag{3.4}$$

当$(\theta',z')\in U$时，对两边取期望并且令$z=z'$，可以得到

$$\mathbb{E}_z\left\|g_\theta(z)-g_{\theta'}(z')\right\|\le\mathbb{E}_z\left[L(\theta,z)\right]\|\theta-\theta'\|\tag{3.5}$$

因此我们可以定义$U_\theta=\{\theta'|(\theta',z)\in U\}$。由于$U$是开集，故$U_\theta$也是开集。因此，根据假设1)，可以定义$L(\theta)=\mathbb{E}_z[L(\theta,z)]$，并且可以得到：

$$\left|W(\mathbb{P}_r,\mathbb{P}_{g_\theta})-W(\mathbb{P}_r,\mathbb{P}_{g_{\theta'}})\right|\le W(\mathbb{P}_{g_\theta},\mathbb{P}_{g_{\theta'}})\le\mathbb{E}_z\left\|g_\theta(z)-g_{\theta'}(z)\right\|\le L(\theta)\|\theta-\theta'\| \tag{3.6}$$

故，对于任意$\theta'\in U_\theta$，$W(\mathbb{P}_r,\mathbb{P}_g)$也服从局部Lipschitz。显然$W(\mathbb{P}_r,\mathbb{P}_g)$处处连续，并且根据Radamacher's theorem它也几乎处处可微。

3) 对于JS散度和KL散度有：

$$JS(\mathbb{P}_0,\mathbb{P}_\theta)=\begin{cases}\log 2 & \theta\ne 0\\ 0 & \theta=0\end{cases}$$

$$KL(\mathbb{P}_0\,\|\,\mathbb{P}_\theta)=KL(\mathbb{P}_\theta\,\|\,\mathbb{P}_0)=\begin{cases}+\infty, & \theta\ne 0\\ 0, & \theta=0\end{cases}$$

当$\theta\to 0$时，JS散度和KL散度均不连续。

证毕。由以上证明可知，W距离在分布$\mathbb{P}_g$和$\mathbb{P}_r$下仍是关于θ的连续函数。因此，通过最小化W距离进行学习对于神经网络来说是有意义的。

推论1 假设g_θ为任意由θ参数化的前馈神经网络，$p(z)$为z上的先验，满足$\mathbb{E}_{z\sim p(z)}[\|z\|]<\infty$ (例如高斯、均匀等)。则假设1)得到满足，因此$W(\mathbb{P}_r,\mathbb{P}_g)$处处连续，且几乎处处可微。

证明 从光滑非线性情况着手，因为g是关于(θ,z)的一阶连续可微函数C^1，那么对于任意固定的(θ,z)，有$L(\theta,z)\le\|\nabla_{\theta,z}g_\theta(z)\|+\varepsilon$是所有$\varepsilon>0$可接受的局部Lipschitz常数。因此，我们只需证明：

$$\mathbb{E}_{z\sim p(z)}\left[\left\|\nabla_{\theta,z}g_\theta(z)\right\|\right]<\infty \tag{3.7}$$

若H是前馈神经网络的层数，有

$$\nabla_{\theta,z}g_\theta(z)=\prod_{k=1}^{H}W_kD_k \tag{3.8}$$

其中，W_k是权重矩阵，D_k是非线性的对角雅可比矩阵。

令$f_{i:j}$是从i层到j层的激活函数，因此有：

$$\nabla_{W_k}g_\theta(z)=\left(\left(\prod_{i=k+1}^{H}W_iD_i\right)D_k\right)f_{1:k-1}(z) \tag{3.9}$$

如果L是非线性的Lipschitz常数，那么有$\|D_i\|\le L$和$\left\|f_{1:k-1}(z)\right\|\le\|z\|L^{k-1}\prod_{i=1}^{k-1}W_i$，把它们整合到一起

$$\begin{aligned}\left\|\nabla_{z,\theta}g_\theta(z)\right\|&\le\left\|\prod_{i=1}^{H}W_iD_i\right\|+\sum_{k=1}^{H}\left\|\left(\left(\prod_{i=k+1}^{H}W_iD_i\right)D_k\right)f_{i:k-1}(z)\right\|\\&\le L^H\prod_{i=1}^{H}\|W_i\|+\sum_{k=1}^{H}\|z\|L^H\left(\prod_{i=1}^{k-1}\|W_i\|\right)\left(\prod_{i=k+1}^{H}\|W_i\|\right)\end{aligned} \tag{3.10}$$

令$C_1(\theta)=L^H\prod_{i=1}^{H}\|W_i\|$和$C_2(\theta)=\sum_{k=1}^{H}L^H\left(\prod_{i=1}^{k-1}\|W_i\|\right)\left(\prod_{i=k+1}^{H}\|W_i\|\right)$，

则有：

$$\mathbb{E}_{z\sim p(z)}\left[\left\|\nabla_{z,\theta}g_\theta(z)\right\|\right]=C_1(\theta)+C_2(\theta)\mathbb{E}_{z\sim p(z)}[\|z\|]<\infty \tag{3.11}$$

证毕。以上定理和推论都表明W距离是比JS散度更加合理的成本函数。接下来进一步证明W距离相较于TV距离、KL散度和JS散度具有最弱的拓扑结构强度。

定理 3.2 设 $\mathbb{P}$ 为紧空间 χ 上的分布，$\left(\mathbb{P}_n\right)_{n\in N}$ 为 χ 上的分布序列。然后，考虑当 $n\to\infty$ 时的所有极限，

1) 以下语句是等效的

$\delta\left(\mathbb{P}_n,\mathbb{P}\right)\to 0$，其中 δ 为总变差距离(TV 距离)；

$JS\left(\mathbb{P}_n,\mathbb{P}\right)\to 0$，其中 JS 为 JS 散度。

2) 以下语句是等效的

$$W\left(\mathbb{P}_n,\mathbb{P}\right)\to 0;$$

$\left(\mathbb{P}_n \xrightarrow{D} \mathbb{P}\right)\to 0$，其中 $\xrightarrow{D}$ 表示随机变量依分布收敛。

3) $KL\left(\mathbb{P}_n \| \mathbb{P}\right)\to 0$ 或者 $KL\left(\mathbb{P} \| \mathbb{P}_n\right)\to 0$ 意味着陈述 1)。

4) 1)中的陈述蕴涵 2)中的陈述。

证明 以下将逐条证明。

1) 首先证明 $\delta\left(\mathbb{P}_n,\mathbb{P}\right)\to 0 \Rightarrow JS\left(\mathbb{P}_n,\mathbb{P}\right)\to 0$。

令 $\mathbb{P}_m=\frac{1}{2}\mathbb{P}_n+\frac{1}{2}\mathbb{P}$（$\mathbb{P}_m$ 取决于 n)是混合分布，根据 TV 距离定义可以证明 $\delta\left(\mathbb{P}_m,\mathbb{P}_n\right)\le\delta\left(\mathbb{P}_n,\mathbb{P}\right)$，且当 $\delta\left(\mathbb{P}_n,\mathbb{P}\right)\to 0$ 时，有 $\delta\left(\mathbb{P}_m,\mathbb{P}_n\right)\to 0$。

取 $f_n=\frac{\mathrm{d}\mathbb{P}_n}{\mathrm{d}\mathbb{P}_m}$ 为 $\mathbb{P}_m$ 和 $\mathbb{P}_n$ 之间的 Radon-Nykodim 导数，通过构造对于每一个 Borel 集 A 有 $\mathbb{P}_n\left(A\right)\le 2\mathbb{P}_m\left(A\right)$，若 $A=\left\{f_n>3\right\}$，可以得到：

$$\mathbb{P}_n\left(A\right)=\int_A f_n\mathrm{d}\mathbb{P}_m\ge 3\mathbb{P}_m\left(A\right) \tag{3.12}$$

于是由 $3\mathbb{P}_m\left(A\right)\le\mathbb{P}_n\left(A\right)\le 2\mathbb{P}_m\left(A\right)$，可得：$\mathbb{P}_m\left(A\right)=0$。

可以通过任意大于 2 的常数得到上述结果，在此我们取常数 3。

取固定的 $\varepsilon>0$，和 $A_n=\left\{f_n>1+\varepsilon\right\}$，于是有：

$$\mathbb{P}_n\left(A_n\right)=\int_{A_n} f_n\mathrm{d}\mathbb{P}_m\ge\left(1+\varepsilon\right)\mathbb{P}_m\left(A_n\right) \tag{3.13}$$

又

$$\begin{aligned}\varepsilon\mathbb{P}_m\left(A_n\right)&\le\mathbb{P}_n\left(A_n\right)-\mathbb{P}_m\left(A_n\right)\\&\le\left|\mathbb{P}_n\left(A_n\right)-\mathbb{P}_m\left(A_n\right)\right|\\&\le\delta\left(\mathbb{P}_m,\mathbb{P}_n\right)\\&\le\delta\left(\mathbb{P}_n,\mathbb{P}\right)\end{aligned} \tag{3.14}$$

故

$$\mathbb{P}_m\left(A_n\right)\le\frac{1}{\varepsilon}\delta\left(\mathbb{P}_n,\mathbb{P}\right) \tag{3.15}$$

此外，

$$\begin{aligned}\mathbb{P}_n\left(A_n\right)&\le\mathbb{P}_m\left(A_n\right)+\left|\mathbb{P}_n\left(A_n\right)-\mathbb{P}_m\left(A_n\right)\right|\\&\le\frac{1}{\varepsilon}\delta\left(\mathbb{P}_n,\mathbb{P}\right)+\delta\left(\mathbb{P}_n,\mathbb{P}_m\right)\\&\le\frac{1}{\varepsilon}\delta\left(\mathbb{P}_n,\mathbb{P}\right)+\delta\left(\mathbb{P}_n,\mathbb{P}\right)\\&\le\left(\frac{1}{\varepsilon}+1\right)\delta\left(\mathbb{P}_n,\mathbb{P}\right)\end{aligned} \tag{3.16}$$

于是根据上述不等式可以得出：

$$\begin{aligned} KL\left(\mathbb{P}_n \| \mathbb{P}_m\right) &= \int \log\left(f_n\right) \mathrm{d}\mathbb{P}_n \\ &\le \log\left(1+\varepsilon\right) + \int_{A_n} \log\left(f_n\right) \mathrm{d}\mathbb{P}_n \\ &\le \log\left(1+\varepsilon\right) + \log 3 \mathbb{P}_n\left(A_n\right) \\ &\le \log\left(1+\varepsilon\right) + \log 3\left(\frac{1}{\varepsilon}+1\right)\delta\left(\mathbb{P}_n, \mathbb{P}\right) \end{aligned} \tag{3.17}$$

对于任意的 $\varepsilon > 0$，对不等式两边取上极限，可得

$$0 \le \limsup KL\left(\mathbb{P}_n \| \mathbb{P}_m\right) \le \log\left(1+\varepsilon\right) + 0 \tag{3.18}$$

即 $KL\left(\mathbb{P}_n \| \mathbb{P}_m\right) \to 0$。

同理，我们可以定义 $g_n = \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{P}_m}$，且令 $B = \{g_n > 3\}$，于是有 $3\mathbb{P}_m\left(B\right) \le \mathbb{P}\left(B\right) \le 2\mathbb{P}_m\left(B\right)$，从而有 $\mathbb{P}_m\left(B\right) = 0$。进一步令 $B_n = \{g_n > 1+\varepsilon\}$ 有：

$$\mathbb{P}\left(B_n\right) = \int_{B_n} g_n \mathrm{d}\mathbb{P}_m \ge \left(1+\varepsilon\right)\mathbb{P}_m\left(B_n\right) \tag{3.19}$$

于是有：

$$\mathbb{P}_m\left(B_n\right) \le \frac{1}{\varepsilon}\delta\left(\mathbb{P}, \mathbb{P}_m\right) \tag{3.20}$$

又

$$\begin{aligned} \mathbb{P}\left(B_n\right) &\le \mathbb{P}_m\left(B_n\right) + \left|\mathbb{P}\left(B_n\right) - \mathbb{P}_m\left(B_n\right)\right| \\ &\le \frac{1}{\varepsilon}\delta\left(\mathbb{P}, \mathbb{P}_m\right) + \delta\left(\mathbb{P}, \mathbb{P}_m\right) \\ &\le \left(\frac{1}{\varepsilon}+1\right)\delta\left(\mathbb{P}, \mathbb{P}_m\right) \end{aligned} \tag{3.21}$$

故且当 $\delta\left(\mathbb{P}, \mathbb{P}_m\right) \to 0$ 时，$\mathbb{P}\left(B_n\right) \to 0$。

根据上述可得：

$$\begin{aligned} KL\left(\mathbb{P} \| \mathbb{P}_m\right) &= \int \log\left(g_n\right) \mathrm{d}\mathbb{P} \\ &\le \log\left(1+\varepsilon\right) + \int_{B_n} \log\left(g_n\right) \mathrm{d}\mathbb{P} \\ &\le \log\left(1+\varepsilon\right) + \log 3 \mathbb{P}\left(B_n\right) \\ &\le \log\left(1+\varepsilon\right) + \log 3\left(\frac{1}{\varepsilon}+1\right)\delta\left(\mathbb{P}, \mathbb{P}_m\right) \end{aligned} \tag{3.22}$$

对两边取上极限得 $0 \le \limsup KL\left(\mathbb{P} \| \mathbb{P}_m\right) \le \log\left(1+\varepsilon\right)$，即 $KL\left(\mathbb{P} \| \mathbb{P}_m\right) \to 0$。

最后，

$$JS\left(\mathbb{P}_n, \mathbb{P}\right) = \frac{1}{2}KL\left(\mathbb{P}_n \| \mathbb{P}_m\right) + \frac{1}{2}KL\left(\mathbb{P} \| \mathbb{P}_m\right) \to 0 \tag{3.23}$$

即证得 $\delta\left(\mathbb{P}_n, \mathbb{P}\right) \to 0 \Rightarrow JS\left(\mathbb{P}_n, \mathbb{P}\right) \to 0$。

接下来证明 $JS\left(\mathbb{P}_n, \mathbb{P}\right) \to 0 \Rightarrow \delta\left(\mathbb{P}_n, \mathbb{P}\right) \to 0$。

根据三角不等和 Pinsker 不等式有

$$\begin{aligned}\delta(\mathbb{P}_n,\mathbb{P}) &\le \delta(\mathbb{P}_n,\mathbb{P}_m)+\delta(\mathbb{P},\mathbb{P}_m)\\ &\le \sqrt{\frac{1}{2}KL(\mathbb{P}_n\|\mathbb{P}_m)}+\sqrt{\frac{1}{2}KL(\mathbb{P}\|\mathbb{P}_m)}\\ &\le 2\sqrt{JS(\mathbb{P}_n,\mathbb{P})}\end{aligned} \tag{3.24}$$

即当 $JS(\mathbb{P}_n,\mathbb{P})\to 0$ 时，有 $\delta(\mathbb{P}_n,\mathbb{P})\to 0$。

2) 根据 W 距离的拓扑性知，如果 $W(\mathbb{P}_n,\mathbb{P})\to 0$，则 $\mathbb{P}_n$ 在弱拓扑意义下收敛到 $\mathbb{P}$。

3) 根据 Pinsker 不等式可得：

$$\begin{aligned}\delta(\mathbb{P}_n,\mathbb{P}) &\le \sqrt{\frac{1}{2}KL(\mathbb{P}_n\|\mathbb{P})}\to 0\\ \delta(\mathbb{P},\mathbb{P}_n) &\le \sqrt{\frac{1}{2}KL(\mathbb{P}\|\mathbb{P}_n)}\to 0\end{aligned} \tag{3.25}$$

故 $KL(\mathbb{P}_n\|\mathbb{P})\to 0$ 或者 $KL(\mathbb{P}\|\mathbb{P}_n)\to 0$ 可以推出陈述 1)。

4) TV 距离诱导是一种强拓扑，意味着当 $\delta(\mathbb{P}_n,\mathbb{P})\to 0$ 时，则 $\mathbb{P}_n$ 几乎处处收敛到 $\mathbb{P}$。而 W 距离诱导了一种比 TV 距离弱的拓扑结构，如果 $W(\mathbb{P}_n,\mathbb{P})\to 0$，则 $\mathbb{P}_n$ 在弱拓扑意义下收敛到 $\mathbb{P}$，适用于支撑集不同的分布。即如果 $\mathbb{P}_n\to\mathbb{P}$ 在 TV 距离下收敛，则在 W 距离下也一定收敛。

由以上定理 3.1、定理 3.2 和推论 1 表明分布 $\mathbb{P}_r$ 和 $\mathbb{P}_g$ 在 W 距离的定义下可以具有良好的连续性、可微性以及收敛性。故在学习低维流形支持的分布时，TV 距离、KL 散度和 JS 散度都不是合理的损失函数，然而 W 距离具有良好的性质，使用 W 距离更为合理。

3.3. Wassertein WGAN (WGAN)

在上一部分我们证明了 W 距离具有良好的理论性质，故选择 W 距离作为衡量分布 $\mathbb{P}_g$ 和 $\mathbb{P}_r$ 之间的差异的指标更为合理，但是直接计算 W 距离的代价太大，因此利用 Kantorovich-Rubinstein 对偶性，将 Wasserstein 距离的计算转化为一个可解的优化问题：

$$W(\mathbb{P}_r,\mathbb{P}_g)=\sup_{\|f\|_L\le 1}\mathbb{E}_{x\sim\mathbb{P}_r}[f(x)]-\mathbb{E}_{x\sim\mathbb{P}_g}[f(x)] \tag{3.26}$$

其中，$\|f\|\le 1$ 表示函数 $f(x)$ 满足 1-Lipschitz 连续。式(3.26)这个形式是 WGAN 的关键，可以利用神经网络来拟合这个 f，即该对偶形式将最优化问题变成了寻找最佳 Lipschitz 函数 f 的问题。

引入模型：

WGAN 中生成器为 $G_\theta(z)$，把噪声 $z\sim\mathbb{P}_z$ 映射到生成样本 $g_\theta(z)\sim\mathbb{P}_g$。

基于对偶形式，将 Wasserstein 距离写为：

$$W(\mathbb{P}_r,\mathbb{P}_g)=\sup_{\|f\|\le 1}\mathbb{E}_{x\sim\mathbb{P}_r}[f_\omega(x)]-\mathbb{E}_{z\sim\mathbb{P}_g}[f_\omega(g_\theta(z))] \tag{3.27}$$

其中，f 是判别器网络 f_ω 的输出，训练 f_ω 以最大化式(3.33)，即估计 Wasserstein 距离。同时训练 $g_\theta(z)$ 来最小化式(3.27)。于是，得到 WGAN 的最终 min-max 目标函数为：

$$\min_\theta\max_{\omega\in\mathcal{W}}\mathbb{E}_{x\sim\mathbb{P}_r}[f_\omega(x)]-\mathbb{E}_{z\sim\mathbb{P}_g}[f_\omega(g_\theta(z))] \tag{3.28}$$

其中，θ 是生成器 G 的参数，ω 是判别器的参数(神经网络的权重)，$\mathcal{W}$ 是满足 1-Lipschitz 条件的判别器参数集合，$f_\omega(x)$ 是判别器的输出。

Kantorovich-Rubinstein 对偶要求 f 是 1-Lipschitz 函数。为了保证这一点，可以训练一个参数化的神经网络，使其权重 ω (神经网络参数)位于紧空间 $\mathcal{W}$ 中。为了使参数 ω 位于紧空间中，可以每次梯度更新

后将权重限制在一个固定的阈值范围内。即采用权重裁剪的方式来近似这个约束：$\omega \leftarrow \text{clip}(\omega,-c,c)$。

本章小结

通过引入 Wassertein 距离，WGAN 在理论上和实践中都显著改善了传统 GANs 训练稳定性和生成质量，这一改进为生成模型的研究和应用开辟了新的方向，具有重要的理论意义和实际价值。

W 距离是连续且可微的(1-Lipschitz 函数)，这意味着可以训练判别器直到达到最优。因为 W 距离处处可微，所以对判别器训练得越多，得到的 Wassertein 梯度就越可靠。

其次，训练判别器直到达到最优，没有发生模式崩溃问题是因为，WGAN 的目标函数使得生成器不会集中在少数几个模式上，而是更全面地学习整个数据分布。而传统 GANs，最优生成器是鉴别器赋予最高值的点的增量之和，主要专注于个别模式。

综上，通过引入 Wassertein 距离，WGAN 有效缓解了 GANs 训练中出现的梯度消失问题和模式崩溃问题，使训练变得更加稳定，生成高质量样本。

4. WGAN-GP 的梯度惩罚常数

4.1. 问题描述

在标准 WGAN 中为了对判别器实施 Lipschitz 约束，使其保证连续性，对判别器的权重矩阵进行了权重裁剪。但是权重裁剪会导致两大问题：弱化模型建模能力，以及梯度爆炸和消失。为了改进以上问题，Gulrajani 等人[7]提出了在目标函数中加入梯度惩罚的替代方法，但是大部分关于加入惩罚项的 GANs 训练文献都专注于实验验证，缺乏理论解释，本章将对其进行数学推导。

4.2. WGAN-GP 和 Lipschitz 约束

为了解决权重裁剪强约束带来的不良行为，现在提出一种实施 Lipschitz 约束的替代方法。当且仅当可微函数的梯度在任何地方都不超过 1 时，它才是 1-Lipschtiz 的，因此直接约束判别器相对于其输入的梯度范数，对随机样本的梯度范数施加惩罚。

具体而言，在原始目标函数中引入以下正则项：

$$L_{GP}=\lambda \mathrm{E}_{\hat{x}\sim\mathbb{P}_{\hat{g}}}\left[\left(\left\|\nabla_{\hat{x}}D\left(\hat{x}\right)\right\|_2-1\right)^2\right] \tag{4.1}$$

其中，x 是从真实样本 $x\sim\mathbb{P}_r$ 和生成样本 $\tilde{x}\sim\mathbb{P}_g$ 之间随机线性插值得到的样本点，即：

$$\hat{x}=\varepsilon x+\left(1-\varepsilon\right)\tilde{x}, \tag{4.2}$$

其中，$\varepsilon\sim U\left[0.1\right]$，$\lambda$ 是惩罚系数。

接下来，将会具体证明梯度惩罚系数为什么取 1。

定理 4.1 令 $\mathbb{P}_r$、$\mathbb{P}_g$ 为紧空间 χ 上的两个分布，则存在一个 1-Lipschitz 函数 f^* 是 $\max_{\|f\|_L\le 1}\mathbb{E}_{y\sim\mathbb{P}_r}\left[f\left(x\right)\right]-\mathbb{E}_{y\sim\mathbb{P}_g}\left[f\left(x\right)\right]$ 的最优解。$\pi\left(x,y\right)$ 是 $\mathbb{P}_r$ 和 $\mathbb{P}_g$ 之间的最优耦合，$\Pi\left(\mathbb{P}_r,\mathbb{P}_g\right)$ 是联合分布 $\pi\left(x,y\right)$ 的集合。那么，如果 f^* 可微，$\pi\left(x=y\right)=0$，$x_t=tx+\left(1-t\right)y$，$0\le t\le 1$，则有：

$$\mathbb{P}_{(x,y)\sim\pi}\left[\nabla f^*\left(x_t\right)=\frac{y-x_t}{\left\|y-x_t\right\|}\right]=1。 \tag{4.3}$$

证明 因为 χ 是紧空间，由[10]知 $\max_{\|f\|_L\le 1}\mathbb{E}_{y\sim P_r}\left[f\left(x\right)\right]-\mathbb{E}_{x\sim P_g}\left[f\left(x\right)\right]$ 存在最优解 f^*，且 $\pi\left(x,y\right)$ 是最优耦合，有

$$\mathbb{P}_{(x,y)\sim\pi}\left[f^*\left(y\right)-f^*\left(x\right)=\left\|y-x\right\|\right]=1 \tag{4.4}$$

令(x,y)满足$f^*(y)-f^*(x)=\|y-x\|$，假设$x\neq y$，这在分布$\pi(x,y)$下发生的概率为1。令$\varphi(t)=f^*(x_t)-f^*(x)$，首先需要证明：

$$\varphi(t)=\|x_t-x\|=t\|y-x\|. \tag{4.5}$$

令t、$t'\in[0,1]$，于是有：

$$|\varphi(t)-\varphi(t')|=|f^*(x_t)-f^*(x_{t'})|\leq\|x_t-x_{t'}\|=|t-t'|\|y-x\| \tag{4.6}$$

因此，φ满足$\|y-x\|$-Lipschitz，从而有：

$$\begin{aligned}\varphi(1)-\varphi(0)&=(\varphi(1)-\varphi(t))+(\varphi(t)-\varphi(0))\\&\leq(1-t)\|y-x\|+t\|y-x\|\\&=\|y-x\|\end{aligned} \tag{4.7}$$

但又因为

$$\varphi(1)-\varphi(0)=f^*(y)-f^*(x)=\|y-x\|, \tag{4.8}$$

故不等式(4.7)变为等式。

特别地，$\varphi(t)-\varphi(0)=t\|y-x\|$，又因为$\varphi(0)=f^*(x)-f^*(x)$，

故

$$\varphi(t)=t\|y-x\|. \tag{4.9}$$

引入变量v，令

$$v=\frac{y-x_t}{\|y-x_t\|}=\frac{y-((1-t)x+ty)}{\|y-((1-t)x+ty)\|}=\frac{y-x}{\|y-x\|} \tag{4.10}$$

根据$f^*(x_t)-f^*(x)=\varphi(t)=t\|y-x\|$，可得

$$f^*(x_t)=f^*(x)+t\|y-x\|, \tag{4.11}$$

接下来，我们对$f^*(x_t)$求偏导：

$$\begin{aligned}\frac{\partial}{\partial v}f^*(x_t)&=\lim_{h\to0}\frac{f^*(x_t+hv)-f^*(x_t)}{h}\\&=\lim_{h\to0}\frac{f^*\left((1-t)x+ty+h\frac{y-x}{\|y-x\|}\right)-f^*(x_t)}{h}\\&=\lim_{h\to0}\frac{f^*\left(x+t(y-x)+h\frac{y-x}{\|y-x\|}\right)-f^*(x_t)}{h}\\&=\lim_{h\to0}\frac{f^*\left(x_{t+\frac{h}{\|y-x\|}}\right)-f^*(x_t)}{h}\\&=\lim_{h\to0}\frac{f^*(x)+\left(t+\frac{h}{\|y-x\|}\right)\|y-x\|-\left(f^*(x)+t\|y-x\|\right)}{h}\\&=\lim_{h\to0}\frac{h}{h}=1\end{aligned} \tag{4.12}$$

如果 f^* 在 x_t 可微，由于 f^* 是 1-Lipschitz 函数，所以 $\left\|\nabla f^*\left(x_t\right)\right\| \leq 1$。

根据简单的 Pythagoras 和单位向量 v，有

$$\begin{aligned} 1 &\leq \left\|\nabla f^*\left(x\right)\right\|^2 \\ &= \left\langle v, f^*\left(x_t\right)\right\rangle^2 + \left\|\nabla f^*\left(x_t\right) - \left\langle v, f^*\left(x_t\right)\right\rangle v\right\|^2 \\ &= \left|\frac{\partial}{\partial v} f^*\left(x_t\right)\right|^2 + \left\|\nabla f^*\left(x_t\right) - v\frac{\partial}{\partial v} f^*\left(x_t\right)\right\|^2 \\ &= 1 + \left\|\nabla f^*\left(x_t\right) - v\right\|^2 \leq 1 \end{aligned} \tag{4.13}$$

由以上不等式结果可得

$$1 = 1 + \left\|\nabla f^*\left(x_t\right) - v\right\|^2, \tag{4.14}$$

于是有 $\left\|\nabla f^*\left(x_t\right) - v\right\|^2 = 0$ 和 $\nabla f^*\left(x_t\right) = v$。进而有：

$$\nabla f^*\left(x_t\right) = \frac{y - x_t}{\left\|y - x_t\right\|} \tag{4.15}$$

故最终得出在联合分布 $\pi(x, y)$ 下 $\mathbb{P}_{(x,y)\sim\pi}\left[\nabla f^*\left(x_t\right) = \frac{y - x_t}{\left\|y - x_t\right\|}\right] = 1$。

由以上证明可得，最优判别器包含连接 $\mathbb{P}_r$ 和 $\mathbb{P}_g$ 耦合点的梯度范数为 1 的直线。故通过对这些插值样本计算判别器输出关于输入的梯度，如果梯度模长偏离 1，就会产生惩罚。

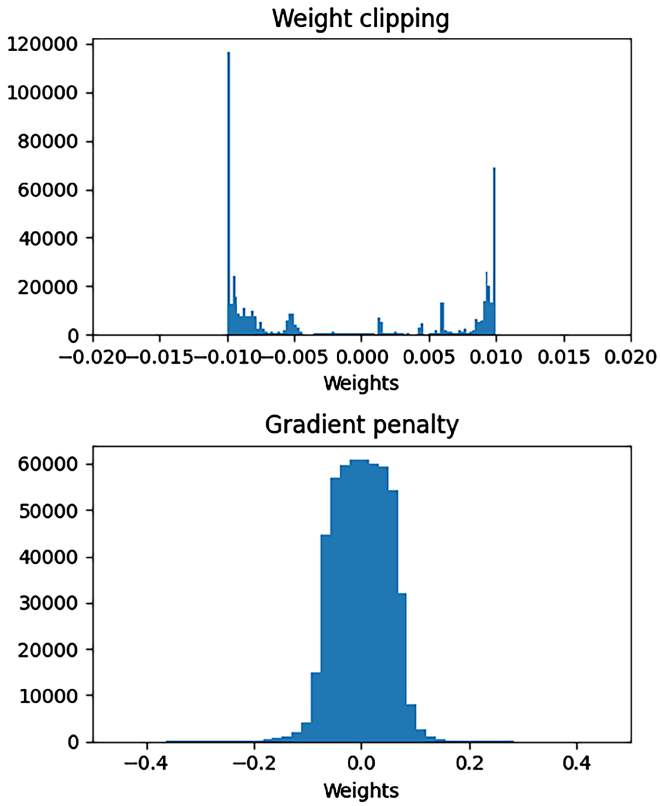

Figure 2. Distribution diagram of discriminator weight under weight clipping and gradient penalty
图 2. 权重裁剪和梯度惩罚下判别器权重分布图

图 2 展示了原始 WGAN 强制对判别器的权重进行裁剪，压制了其表达能力，使权重朝向两个极端值靠近。而梯度惩罚则呈现出标准高斯形状，分布更自然，使网络能更好训练。

本章小结

梯度惩罚是一种修改判别器目标函数的方法，通过在目标函数中加入梯度惩罚项来避免梯度消失或者梯度爆炸的问题。这种方法相对于权重裁剪更平滑、更具鲁棒性，能够有效提升训练的稳定性和生成图像的质量。同时，由于梯度惩罚不限制网络权重的具体取值，使得可以使用更深、更复杂的网络结构，从而增强模型的表达能力。

5. 研究总结

本论文围绕生成对抗网络(GAN)训练过程中的不稳定性问题，展开了系统的理论分析。首先，从传统 GAN 的目标函数出发，揭示了其在使用标准损失函数时所面临的梯度消失问题。

其次，针对这一问题，论文对基于 Wasserstein 距离作为 GANs 训练的目标函数，系统地证明了其连续性、可微性与收敛性，使得生成器在判别器达到最优时仍然能够获得有效梯度，显著提升训练的稳定性和收敛性。进而在分析 WGAN 训练策略的基础上，对于引入梯度惩罚以取代传统的权重裁剪，从数学上严格推导了其满足 Lipschitz 约束的理论依据。

参考文献

[1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., *et al*. (2014) Generative Adversarial Nets. *Advance in Neural Information Processing Systems*, **27**, 2672-2680.

[2] Karras, T., Aila, T., Laine, S. and Lehtinen, J. (2017) Progressive Growing of GANs for Improved Quality, Stability, and Variation. *Proceeding of the Advance in Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 1-26.

[3] Brock, A., Donahue, J. and Simonyan, K. (2018) Large Scale GAN Training for High Fidelity Natural Image Synthesis. *Proceeding of the Advance in Neural Information Processing Systems*, Montréal, 3-8 December 2018, 1-35.

[4] Arjovsky, M. and Bottou, L. (2017) Towards Principled Methods for Training Generative Adversarial Networks. *International Conference on Learning Representations*, Toulon, 24-26 April 2017, 1-10.

[5] Arjovsky, M., Chintala, S. and Bottou, L. (2017) Wassertein Generative Adversarial Networks. *International Conference on Machine Learning*, Sydney, 6-11 August 2017, 214-223.

[6] Mescheder, L., Geiger, A. and Nowozin, S. (2017) Which Training Methods for GANs Do Actually Converge. *Proceeding of the Advance in Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 3481-3490.

[7] Gulrajani, I., Ahmed, F., Arjovsky, M., *et al*. (2017) Improved Training of Wasserstein GANs. *Advance in Neural Information Processing Systems*, **30**, 1-11.

[8] Miyato, T., Kataoka, T., Koyama, M. and Yoshida, Y. (2018) Spectral Normalization for Generative Adversarial Networks. *Proceeding of the Advance in Neural Information Processing Systems*, Montréal, 3-8 December 2018, 1-26.

[9] Zhang, H., Goodfellow, I., Metaxas, D. and Odena, A. (2019) Self-Attention Generative Adversarial Networks. *Proceedings of the* 36*th International Conference on Machine Learning*, California, 9-15 June 2019, 7354-7363.

[10] Maddison, C.J., Mnih, A. and The, Y.W. (2016) The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. *International Conference on Machine Learning*, New York, 19-24 June 2016, 2951-2960.