TOWARDS PRINCIPLED METHODS FOR TRAINING GENERATIVE ADVERSARIAL NETWORKS

Martin Arjovsky Courant Institute of Mathematical Sciences martinarjovsky@gmail.com

Leon Bottou Facebook AI Research

leonb@fb.com

本文共分为三个部分。第一部分介绍了当前问题。第二部分对生成性对抗网络训练中的不稳定性和饱和性等（第一部分提出的问题）进行了严格的研究和论证。第三部分探讨了解决这些问题的时间和理论指导，同时介绍了研究这些问题的新工具。

# 第一部分：

1、GAN在生成图像方面取得了巨大的成功。但是不管怎样他们都非常难以训练（训练困难、生成器和判别器的loss无法指示训练进程、生成样本缺乏多样性等问题），目前大多数的论文都致力于从启发的角度寻找稳定的体系结构，但是几乎没有任何理论可以Gan训练的不稳定行为。这是要改变的一个现状。

2、以前的生成模型训练依赖于极大似然或者最小化KL散度，

这个代价函数有一个优点，那就是在Pg=Pr时有唯一的最小值

如果两个分布相等，则KL散度达到最小值。但是KL散度是不对称的，论文中是以为KL(Pr||Pg)例说明的:

如果Pr>Pg，尤其是Pr>0且Pg→0，那么散度变得无穷大，说明如果生成模型没能生成接近真实数据的数据，cost会很大；相反如果Pr<Pg,尤其是Pr→0且Pg>0时，散度趋于零，说明如果生成模型生成假数据，cost会很小，这就意味着我们的模型会以极低的成本生产假样本，但是这不是我们想要的结果。

用JS散度替代KL(Pg||Pr)，同样没有解决问题（梯度消失，后面解释）。



3、Gan的目标函数是：

最优的判别器长这样：

（看成一个样本x来自真实分布和生成分布的可能性的相对比例）

带入目标函数后：

也就是说，优化Gan的目标函数，就是最小化两个分布的JS散度，因此我们首先应尽可能先训练鉴别器到最优（所以θ上的函数更接近于JSD，因为目标函数变成了最优鉴别器的JSD），然后再在θ上gradient steps，两个步骤交替进行。但是实践中发现，当鉴别器更好时，生成器的更新就会 持续恶化，这就是GAN之所以训练困难的所在。

# 第二部分

1.直观上理解：

JS散度的问题：如果两个分布不重叠或者重叠部分可忽略不计，那么其JS散度达到最大值且恒为常数，这样的话梯度消失，生成模型难以训练。

那么什么算重叠部分可忽略呢？等价于说两个分布是高维空间的低维流型（不是全维），那么他们就撑不满高维空间，也就是说他们在高维空间不连续，也就是说即使有重叠部分，但是在高维空间测度为0（比如说两条曲线相交于一点，这一点长度为0，长度为二维空间的测度）。

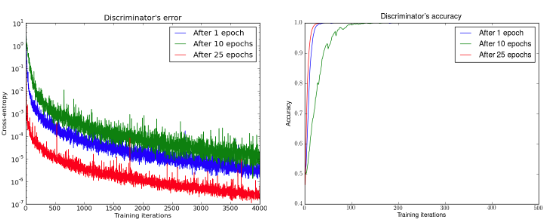
2.引入数学定理分析

**引理1**：已知g:Z→X，即Z经过一系列变换成X，g(z)就可以说是一个不超过Z的维度的低维流型，如果X的维度比Z高，那么g(z)在X上测度为0.

## 2.1完全判别定理：

**定理2.1**：如果Pr和Pg有不相交的紧凑的支撑集，则一定有一个最优判别器D※可以将两个分布分开。

然后做了个实验：



将三个DCGAN分别训练1、10、25个epoch，然后固定他们的生成器去从头训练判别器（当然也是三个）。从左图我们可以看到，判别器的error（交叉熵）在几千次迭代内迅速的下降到0，这时的样本（我认为说的是生成样本）非常好，而且两个分布的支撑集有可能相交。右图显示的是鉴别器的准确性，他在很少的迭代次数（有时不到50次）就收敛到了1。注意y轴是对数坐标轴。

这个实验放在这很突兀，上下文也没什么联系，不知道作者想说什么。我猜测作者是想用这个实验做为辅证，证明2.1完全判别定理分割支撑集的作用，说明存在最优判别器可以精准分割。

**定义2.1：**定义transversality（横截），M and P 是Rd的子流形，令x∈M∩P是两个流形的交点，如果，我们说M and P 在x上横截性相交（intersect transversally）。其中TXM表示x周围M的切线空间。（假装理解了，反正后面没再提）

**定义2.2: 完全对齐的定义，**We say that two manifolds without boundary M and P perfectly align if there is an x∈M∩P such that M and P don’t intersect transversally in x.

如果存在一个x∈M∩P使得M和P没有横截性相交，我们说两个没有边界的流形M和P没有完全对齐。

有趣的是，我们可以在实践中安全的假设任何两个流形都不会完全对齐。这是可以做到的，因为两个流形上的任意小的随机扰动将导致他们横向相交或者根本不相交。

**引理2：**M and P 是Rd的子流形，是 ，连续随机向量，且 ，则

**引理3：**M和P不完全对齐，且不是全维，使 ，如果M和P没有边界，则ℒ还是一个流形，且他的维度严格小于M和P的维度。如果他们有边界，则ℒ是最多四个严格低维流形的并集。不管怎样ℒ在M、P上的测度都为0.

**定理2.2：**

假设Pr和Pg是两个分布，他们的支撑集是M和P，M、P是不完全对齐且不是全维的，进一步假设Pr和Pg在他们的流形中是连续的，这意味着如果有一个集合A在M中度量为0，那么Pr（A）= 0（Pg一样）。然后存在一个精度为1的最优判别器D\*：x→[0,1]，

D\*在x的邻域中光滑并且 。

**定理2.3：**假设Pr和Pg是两个分布，他们的支撑集是M和P。M、P是不完全对齐且不是全维的，进一步假设Pr和Pg在他们的流形中是连续的。在如上所述的情况下，JS散度为常数log2，KL散度为无穷大。

这里指出，即使两个流形任意靠近，他们的散度也会达到最大值，即使生成器生成的样本看起来非常好，但是两个KL散度都是无穷大的，也就是说，如果两个散度都是无穷大的那么用梯度下降来最小化他们是不可能的。

## 2.2各成本函数的后果以及存在的问题

接下来讨论的都是生成器的训练问题。

这里分为了两个部分，第一部分讲的是原始GAN的问题，第二部分讲的是改进后的GAN的问题。

原始GAN就是尽可能的把真实样本分为正利，生成样本分为负例：



对于生成器一开始提出的原始损失函数是：

后来又紧接着提出了一个改进版的：

### 2.2.1 原始GAN的问题

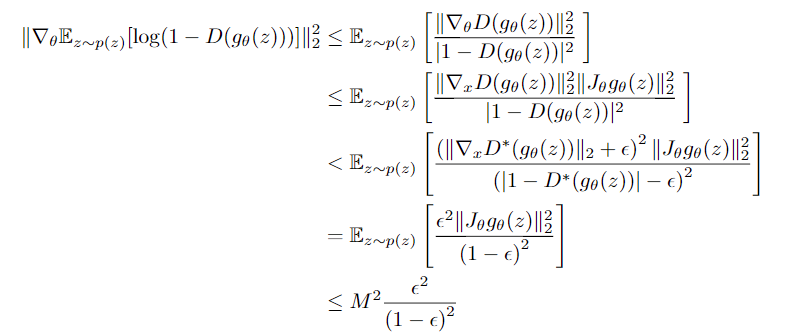
现在我们看看通过鉴别器将梯度传给生成器会发生什么。

**定理2.4：**令Pg是生成分布，Pr为真是分布，如果满足定理2.1和2.2，满足 and  ，

则，

由此可知， 。

证明：



一句话概括就是：判别器训练的越好，生成器梯度消失的越严重。

我们从另一个角度理解论文，由第一部分我们可以知道目标函数可以写成JS散度的形式： 。我们最小化loss也就是最小化两个分布之间的JS散度，然而问题就出在这里。我们最初的目的是让两个分布的JS散度尽可能的小，从而生成的数据才可以以假乱真，但是这必须在两个分布具有高重叠率的情况下才成立的，但是如果两个分布完全没有重叠的部分或者重叠的部分可以忽略不计，那么他们的JS散度就会变成log2。因为对于任意一个x都只有四种可能：



KL散度:

JS散度：

第一种情况说明两个分布毫无重叠，对计算JS散度无贡献，第二种情况重叠部分可以忽略不计，所以贡献也为0，第三第四带入公式得到log2，然而log2是一个常数这对于梯度下降来说意味着梯度为0，由此生成器学习不到任何的东西，即使判别器是一个接近最优的D→D\*的判别器，那么生成器也很有可能会梯度消失。

我们刚刚假设的是Pr和Pg毫无重叠或者重叠部分忽略不计的情况，那么出现这种情况的概率有多大？答案是相当大。知乎上给出的比较严谨的答案是：**当Pr与Pg的支撑集是高维空间中的低维流形时，Pr和Pg重叠部分测度为0的概率为1。**

注，解释几个名词：

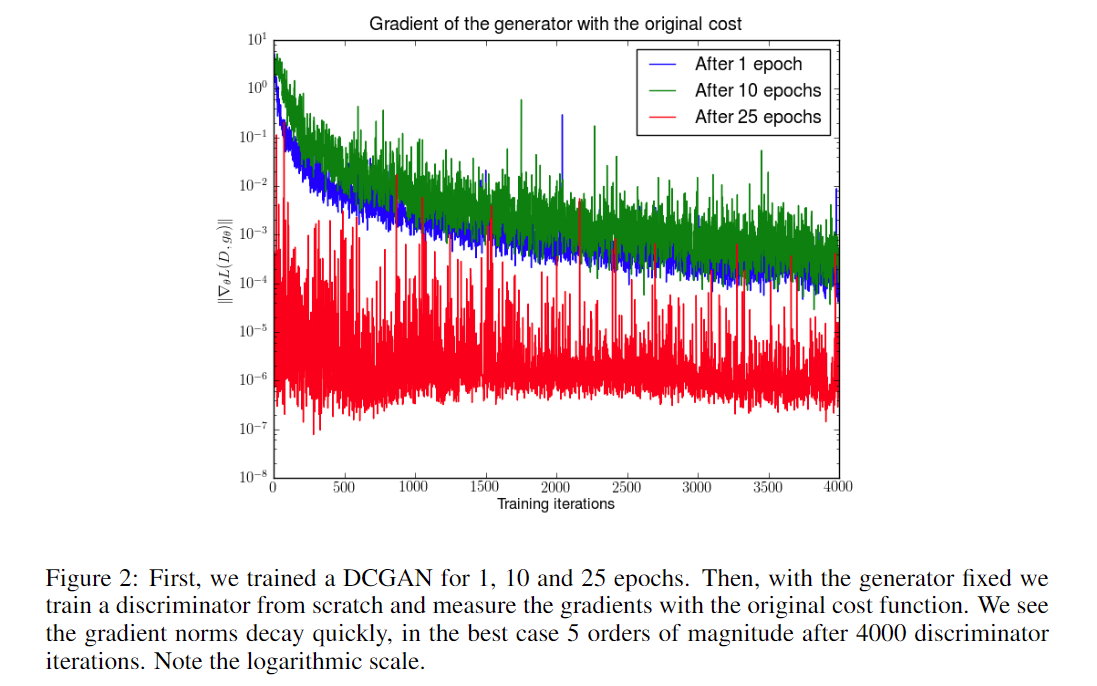
支撑集（support）其实就是函数的非零部分子集，比如ReLU函数的支撑集就是(0, +∞)，一个概率分布的支撑集就是所有概率密度非零部分的集合。

流形（manifold）是高维空间中曲线、曲面概念的拓广，我们可以在低维上直观理解这个概念，比如我们说三维空间中的一个曲面是一个二维流形，因为它的本质维度（intrinsic dimension）只有2，一个点在这个二维流形上移动只有两个方向的自由度。同理，三维空间或者二维空间中的一条曲线都是一个一维流形。

测度（measure）是高维空间中长度、面积、体积概念的拓广，可以理解为“超体积”。所谓测度为0，也就是上面说的重叠部分忽略不计的意思。

由此，我们得出了第一个论证，在最优或者近似最优判别器下，最小化生成器的loss等价于最小化Pr和Pg之间的JS散度，而由于Pr与Pg几乎没有不可忽略的重叠，所以无论他们相距多远JS散度都是常数log2，最终导致生成器的梯度近似为0，梯度消失。有了上面两个方面的分析，我们得知原始GAN的不稳定原因是因为判别器训练的火候难以把握，太好太坏都不行，所以原始Gan才难以训练。

然后作者又做了个实验，作为辅证：



先分别将DCGAN训练1，20，25个epoch，然后固定生成器不动，判别器重新随机初始化从头开始训练，对于原始Gan形式的生成器loss产生的梯度可以打印出其尺度的变化曲线，可以看到随着判别器的训练，生成器的梯度均迅速衰减。注意y轴是对数坐标轴。

### 2.2.2 第二种Gan形式存在的问题

第二种Gan（生成器）形式：

我们已经知道，在最优判别器 的情况下：





然后我们把KL散度也变换成带有的形式：



然后我们将上述几个式子综合科得：



注意，我们训练生成器是需要固定住判别器的，所以最有一项可以视为常数，所以最小化这个损失函数就相当于最小化 ，问题就出现了……

第一点：我们来看一下这个式子，他是一个KL散度减去两个JS散度，也就是需要同时最小化KL散度又要最大化JS散度，一个要拉近距离一个又要推远距离，这会严重导致梯度上的不稳定。

第二点：前面第一部分我们提到了，KL散度是一个不对称的度量，KL(Pg||Pr)和KL(Pr||Pg)是不一样的（一般Gan网络中最小化的是KL(Pr||Pg)，而论文分析中会分析最小化KL(Pg||Pr)）。

KL散度是不对称的，以KL(Pr||Pg)为例:

如果Pr>Pg，那么x是来自真实数据的概率高于生成数据的点，尤其是Pr>0且Pg→0，那么散度变得无穷大，说明如果生成模型没能生成真实数据，cost会很大；相反如果Pr<Pg,尤其是Pr→0且Pg>0时，散度趋于零，说明如果生成模型生成假数据，cost会很小，这就意味着我们的模型会以极低的成本生产假样本，如此生成器宁可多生成一些重复但是很“安全”的样本，也不愿意去生成多样性的样本，这就是通常所说的模式崩塌的核心。

原始GAN问题的根源可以归结为两点，一是等价优化的距离衡量（KL散度、JS散度）不合理，二是生成器随机初始化后的生成分布很难与真实分布有不可忽略的重叠。

# 第三部分

现在的一个重要的问题就是解决Gan的训练不稳定性和梯度渐变消失的问题。为了打破这些定理的假设，我们可以给鉴别器的输入加上连续的噪声，从而平滑概率质量的分布。

在第三部分中，作者对于两个分布没有重叠的问题给出了一个过渡的解决方案，那就是在两个样本分布中各加入一个噪声ε（变成了 ），强行让这两个分布发生不可忽略的重叠，而一旦有了重叠，那么JS散度也就可以真正的发挥作用，从而使两个分布靠近，使他们重叠的部分越来越多，如此梯度消失也就解决了。训练时，慢慢的对两个分布模拟退火，慢慢的减少方差，直到两个低维流形“本体”有所重叠时，把噪声拿掉，JS还是可以继续发挥作用拉近两个低维流形，直到他们接近重合。

但是，两个加上噪声的分布的JS散度虽然可以在某种程度上代表两个原本分布的距离，但是因为加噪的JS散度的具体数值受到噪声的方差影响，随着噪声的退火，前后的数值就没有办法比较了，所以他并不能成为Pr和Pg距离的本质性衡量，进而也就不可以通过由JS散度构成的最优判别器的loss反应训练过程。

于是，作者就提出了用Wasserstein距离代替JS散度，解决了稳定训练和进程指标的问题。