本周推送的话题是WGAN——Wasserstein GAN[2]，这篇文章于2017年1月26日出现在arXiv上，并迅速得到了大家的热议，在reddit上有专门的帖子讨论这篇文章，甚至连Ian Goodfellow，GAN的提出者也参与了讨论。文末有reddit讨论区的地址。感谢郑华滨同学推荐的WGAN前传[1]：Towards Principled Methods for Training Generative Adversarial Networks，Martin Arjovsky把文章投给了ICLR 2017，前天公布的录取结果显示文章中了oral。前传是GAN的理论研究，WGAN不仅在理论上做了进一步研究，还提出了针对GAN一系列问题的解决方案，包括mode collapse、训练不稳定等。

两篇文章的篇幅都比较长，而且大部分内容是理论证明，需要用到拓扑、概率以及分析的一些知识，笔者的拓扑只学了点集拓扑，流形的拓扑纯属空白，一边补知识一边看论文。本期推送介绍WGAN两篇论文的每个引理、定理、推论都讲了些什么，以及WGAN与general GAN有什么不同，不涉及具体证明。通俗易懂地解读WGAN的工作郑华滨同学已经做了，这方面我写得没他好，推荐大家读他的专栏文章：

<https://zhuanlan.zhihu.com/p/25071913>

下面进入正题。

PS: 本次推送文章特别长！在微信中浏览效果并不好，点击阅读原文提取本次推送的PDF版本。

**GAN回顾**

我们来回顾一下，GAN的loss function:



这个loss其实就是交叉熵。对于discriminator（D），它的任务是最小化 -L(G,D) ，即



如果采用零和博弈，generator（G）的目标是最小化L(G,D)，而实际操作发现零和博弈训练效果并不好，G的目标一般采用最小化



需要注意的是，一般来说，更新D是，G是固定的；更新G时，D是固定的，因此，对于零和博弈的情形，G的目标实际上只跟第二项有关，而与第一项无关（与真实数据相关），即



Martin 称这个loss为original cost function（参见[1] 2.2.1章节），而实际操作中采用的loss为the –log D cost（参见[1] 2.2.2章节）。

**GAN存在的问题：初探**

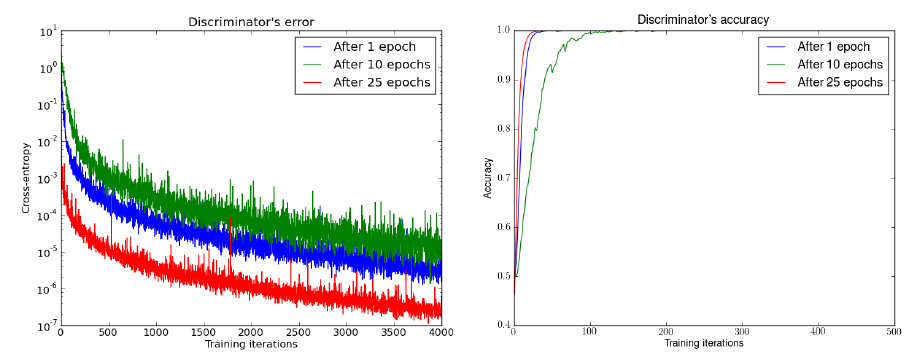
Goodfellow证明了，当G固定时，D的loss具有上界



其中



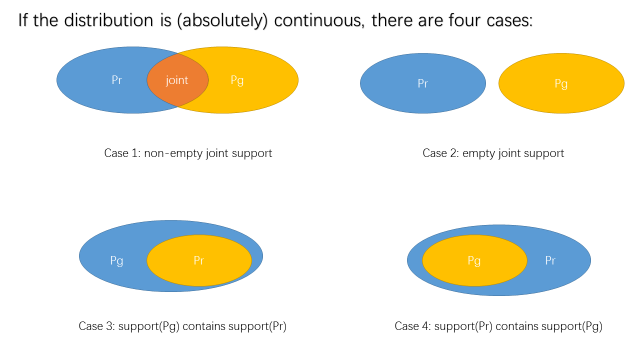
当固定G时，训练D直到收敛，可以发现D的loss会越来越小，趋于0，这表明JSD(Pr || Pg)被最大化了，并且趋于log2。如下图所示。



而这会导致什么问题呢？在实践中人们发现，当D训练得更精确，G的更新会变得越差，训练变得异常地不稳定。为什么会产生这些这样的问题？之前一直没有人给出解答。

JSD(**P**r || **P**g)达到最大化，有两种可能：

* 概率分布不是（绝对）连续的，也就是说，它没有密度函数。我们常见的分布一般都有密度函数。如果概率分布是定义在一个低维的流形上（维度低于全空间），那它就不是绝对连续的。
* 分布是绝对连续的，但是两者的支撑集没有交集。两个分布的支撑集不外乎包含以下四种情形：



经过计算可以发现（参见下期推送），case1 的JSD小于log2，case2的JSD等于log2，case3和4的JSD也不超过log2。实际上这很好理解，两个分布差异越大，交叉熵越大。

这是比较直观的解释，更进一步地，作者从理论上进行了严格的分析和证明。

**GAN存在的问题：理论分析**

**Lemma 1**：设



是一个由仿射变换和逐点定义的非线性函数（ReLU、leaky ReLU或者诸如sigmoid、tanh、softplus之类的光滑严格递增函数）复合得到的复合函数，则g(Z)包含在可数多个流形的并集中，并且它的维数至多为dim(Z)。因此，若dim(Z) < dim(X)，则g(Z)在X中测度为0。

Lemma1表明，若generator (G)是一个神经网络，并且G的输入（随机高斯噪声）的维数比产生的图像的维数低，则无论怎样训练，G也只能产生整个图像空间中很小的部分，有多小呢？它在图像空间中只是一个零测集。零测集是什么概念呢，举些例子，二维空间中的一条直线、三维空间中的一个平面。二维平面在三维空间中是没有“厚度”的，它的体积是0。

我们训练GAN时，训练集总归是有限的，**P**r一般可以看成是低维的流形；如果**P**g也是低维流形，或者它与**P**r的支撑集没有交集，则在discriminator (D)达到最优时，JSD就会被最大化。D达到最优将导致G的梯度变得异常地差，训练将变得异常不稳定。

下面的几个定理、引理就是在证明**P**g在上述两种情况下，最优的D是存在的。

**Theorem 2.1**: 若分布**P**r和**P**g的支撑集分别包含在两个无交紧致子集M和P中，则存在一个光滑的最优discriminator D\*: X 🡪 [0,1]，它的精度是1，并且，对任意的



有



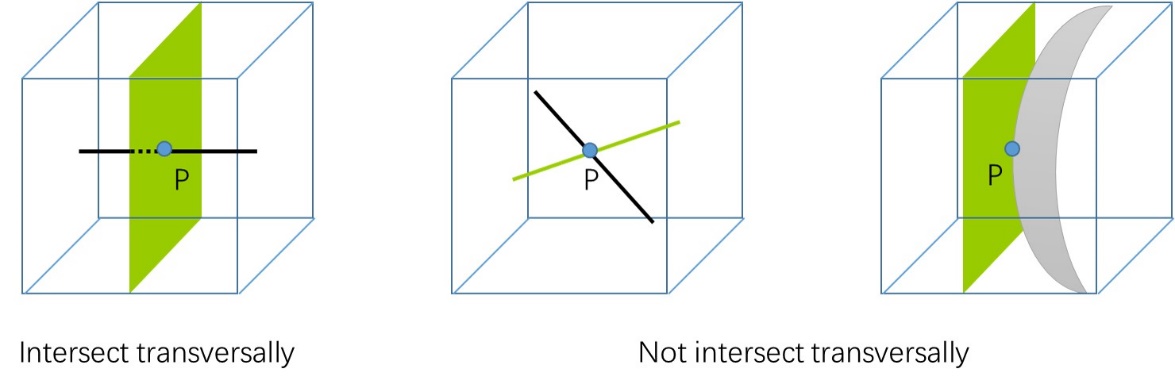
定理2.1是什么意思呢？如果两个概率分布的支撑集没有交集，则完美的D总是存在的，并且（在两个分布的支撑集的并集上）D的梯度为0，也就是说，这时候任何梯度算法都将失效。这就是GAN训练的时候，（在两个概率分布的支撑集没有交集的情况下）当D训练效果很好的时候，G的更新将变得很差的原因。

**Lemma 2**: 设M和P是R^d的两个非满维度的正则子流形，再设η 和 η’ 是任意的两个独立连续随机变量，定义两个扰动流形M’ = M + η，P’ = P + η’，则



Lemma 2是为定理2.2做准备，它表明任意两个非满维的正则子流形都可以通过微小的扰动使得它们不是完美对齐（not perfectly align）的，即它们的交点都是横截相交（intersect transversally）的。横截相交和完美对齐的严谨定义将在下期推送中给出，在这里形象地说明一下：

**横截相交（intersect transversally）**：对两个子流形，在任意一个交点处，两者的切平面能够生成整个空间，则称两个子流形横截相交。当然，如果它们没有交集，则根据定义，它们天然就是横截相交的。下图给出了一个示例，在交点P处，平面的切平面是其自身，直线的切平面也是其自身，它们可以张成全空间，因此是横截相交的，而两个直线没办法张成全空间，因此不是横截相交的；如果两个流形是相切的，在切点处它们的切平面是相同的，也不可能张成全空间，因此也不是横截相交的。



**完美对齐（perfectly align）**: 如果两个子流形有交集，并且在某个交点处，它们不是横截相交的。

**P**r和**P**g的支撑集如果存在交集，那么根据lemma2，我们总可以通过微小的扰动使得它们不是完美对齐的，也就是说，它们是横截相交的。

**Lemma 3**： 设M和P是R^d的两个非完美对齐，非满维的正则子流形，L=M∩P，若M和P无界，则L也是一个流形，且维数严格低于M和P的维数。若它们有界，则L是至多四个维数严格低于全空间的流形之并。无论哪种情形，L在M或者P中的测度均为0。

Lemma 3说的是，两个正则子流形（满足一定条件：非完美对齐，非满维）的交集的维数要严格低于它们自身的维数。也就是说，它们的交集只是冰山一角，小到相对它们自身可以忽略。对于**P**r和**P**g的支撑集来说，根据Lemma 2，我们总可以通过微小扰动使得它们是非完美对齐的，在根据Lemma 3，**P**r和**P**g的交集是微不足道。

**Theorem 2.2**: 设**P**r和**P**g分别是支撑集包含在闭流形M和P中的两个分布，且它们非完美对齐、非满维。进一步地，我们假设**P**r和**P**g在各自的流形中分别连续，即：若集合A在M中测度为0，则**P**r(A) = 0（在**P**g上也有类似结果）。则存在精度为1的最优discriminator D\*: X 🡪 [0,1]，并且几乎对M 或者P中的任意点x，D\*在x的任意邻域内总是光滑的，且有



定理2.1证明的是对于**P**r和**P**g无交的情形下，最优的discriminator是存在的。定理2.2承接Lemma 3，它证明了在**P**r和**P**g的支撑集有交集，且横截相交的情况下，最优的discriminator是存在的。这两个定理实际上把两种可能导致D最优，且梯度消失的情形在理论上做出证明，由于梯度的消失，G的更新将得不到足够的梯度，导致G很差。

**Theorem 2.3**: 在定理2.2的条件下，有



定理2.3表明，随着D越来越好，D的loss将越来越小，趋于0，因此**P**r和**P**g的JSD被最大化，达到最大值log2，这时，**P**r和**P**g的交叉熵达到无穷大，也就是说，即使两个分布之间的差异任意地小，它们之间的KL散度仍然会被最大化，趋于无穷。这是什么意思呢？利用KL散度来衡量分布间的相似性在这里并不是一个好的选择。因此，我们有必要寻求一个更好的衡量指标。

**Theorem 2.4**: 设



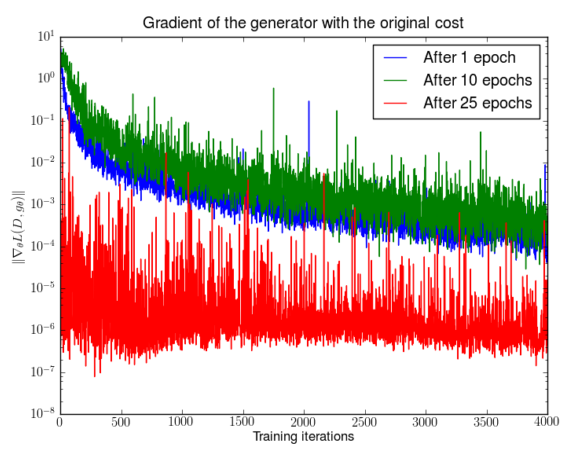
为一个可微函数，由它导出分布**P**g，再设**P**r为真实数据的分布，D是一个可微的discriminator。若定理2.1和定理2.2的条件能被满足，且



则



定理2.4 探究了generator在前面所述情况下回出现什么问题，它从理论上给出了，若G采用original cost function（零和博弈），那么它的梯度的上界被D与最优的D\*之间的距离bound住。说人话就是，我们训练GAN的时候，D越接近最优的D\*，则G的梯度就越小，如果梯度太小了，梯度算法不能引导G变得更好。下图给出了G的梯度变化（固定G，训练D），注意到随着训练的进行，D将变得越来越精确，这时G的梯度强度将变得越来越小，与理论分析符合。



**Corollary 2.1**: 在定理2.4的假定下，有



推论2.1是定理2.4的极限情况。

**Theorem 2.5**: 设



再设固定θ = θ\_0时，最优的discriminator为



则



定理2.5研究了G的loss为the –logD alternative时将会出现的问题。我们可以看到，当JSD越大时，G的梯度反而会越小，也就是说，它可能会引导两个分布往相异的方向，此外，上式的KL项虽对产生无意义图像会有很大的惩罚，但是对mode collapse惩罚很小，也就是说，GAN训练时很容易落入局部最优，产生mode collapse。KL散度不是对称的，但JSD是对称的，因此JSD并不能改变这种状况。这就是我们在训练GAN时经常出现mode collapse的原因。

**Theorem 2.6**: 设g\_θ: Z 🡪 X是一个可微函数，由它可导出分布**P**g，再设**P**r为真实数据的分布，且满足定理2.1或者2.2之一的条件。令D为一个discriminator，满足



为高斯白噪声，且



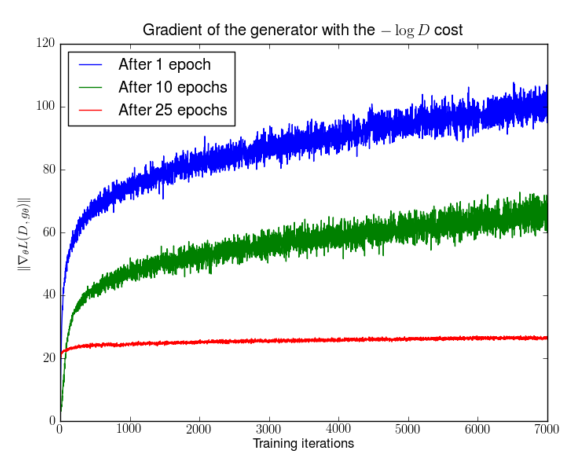
也为高斯白噪声。则



的每一维均服从期望和方差为正无穷的中心化柯西分布。

定理2.6告诉我们，若G采用the –logD cost，在定理2.1或者2.2的条件下，当D与D\*足够接近时，G的梯度会呈现强烈震荡，这也就是说，G的更新会变得很差，可能导致GAN训练不稳定。

下图给出了定理2.6的实验模拟的效果，在DCGAN尚未收敛时，固定G，训练D将导致G的梯度产生强烈震荡。当DCGAN收敛时，这种震荡得到有效的抑制。



既然general GAN采用的loss不是一种好的选择，有什么loss能够有效避免这种情形吗？

**一个临时解决方案**

一个可行的方案是打破定理的条件，给D的输入添加噪声。后续的几个定理对此作了回答。

**Theorem 3.1**: 若X满足分布**P**\_X，且它的支撑集落在M中，ε是一个密度函数为P\_ε的绝对连续分布，则

****

也是一个绝对连续分布，具有密度函数



**Corollary 3.1**: 若



则



若



则



若



则



定理3.1和推论3.1表明，ε的分布会影响我们对距离的选择。

**Theorem 3.2**: 设**P**r和**P**g分别是支撑集落在M和P中的两个分布，



则G的梯度具有以下形式



其中a(z)、b(z)是两个正值函数。更进一步地，



定理3.2证明了G的梯度可以分为两项，第一项表明，G会被引导向真实数据分布移动，第二项表明，G会被引导向概率很高的生成样本远离。作者指出，上述的梯度格式具有一个很严重的问题，那就是由于g(Z)是零测集，D在优化时将忽略该集合；然而G却只在该集合上进行优化。进一步地，这将导致D极度容易受到生成样本的影响，产生没有意义的样本。

**Corollary 3.2**: 设



则



对D的输入添加噪声，在训练的过程中将引导噪声样本向真实数据流形的方向移动，可以看成是引导样本的一个小邻域向真实数据移动。这可以解决D极度容易受到生成样本的影响的问题。

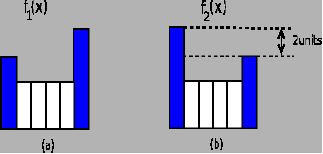
**Wasserstein距离**

**定义3.1**: X上的两个分布P、Q的Wasserstein度量W(P,Q)定义为



其中，Г是X×X上所有具有边界分布P和Q的联合分布集。

Wasserstein距离通常也称为转移度量或者EM距离（地动距离），它表示从一个分布转移成另一个分布所需的最小代价。下图给出了一个离散分布下的例子，将f1(x)迁移成f2(x)最小代价即是移动f1(x)在最大值处的2个单位的概率到最小值处，这样就得到了分布f2(x)。更复杂的离散情形需要通过求解规划问题。



**Lemma 4**: 若 ε 是零均值的随机向量，则



其中，



是ε的方差。

引理4表明，一个分布与它添加扰动之后的分布的Wasserstein距离能被扰动的标准差bound住。

**Theorem 3.3**: 设**P**r和**P**g是任意两个分布，ε 是一个零均值，方差为V的随机向量。若



的支撑集落在直径为C的球内，则



定理3.3告诉我们一个有趣的事实，上式右边两项均能被控制。第一项可以通过逐步减小噪声来逐步减小；第二项可以通过训练GAN（给D的输入添加噪声）来最小化。

作者指出，这种通过给D的输入添加噪声的解决方案具有一大好处，那就是我们不需要再担心训练过程。由于引入了噪声，我们可以训练D直到最优而不会遇到G的梯度消失或者训练不稳定的问题，此时G的梯度可以通过推论3.2给出。

总而言之，WGAN的前传从理论上研究了GAN训练过程中经常出现的两大问题：G的梯度消失、训练不稳定。并且提出了利用地动距离来衡量**P**r和**P**g的相似性、对D的输入引入噪声来解决GAN的两大问题，作者证明了地动距离具有上界，并且上界可以通过有效的措施逐步减小。

这可以说是一个临时性的解决方案，作者甚至没有给出实验进行验证。在WGAN[2]这篇文章中，作者提出了更完善的解决方案，并且做了实验进行验证。下面我们就来看一下这篇文章。

**Wasserstein GAN**

**常见距离**

Martin Arjovsky在WGAN论文进一步论述了为什么选择Wasserstein距离（地动距离）。

设X是一个紧致度量空间，我们这里讨论的图像空间（[0,1]^d）就是紧致度量空间。用Σ表示X上的所有博雷尔集，用Prob(X)表示定义在X上的概率度量空间。给定Prob(X)上的两个分布**P**r, **P**g，我们可以定义它们的距离/散度（请注意：散度不是距离，它不是对称的。距离和散度都可以用于衡量两个分布的相似程度）：

1. 全变差距离（TV距离）



1. KL散度



1. JS散度



1. Wasserstein距离（地动距离）



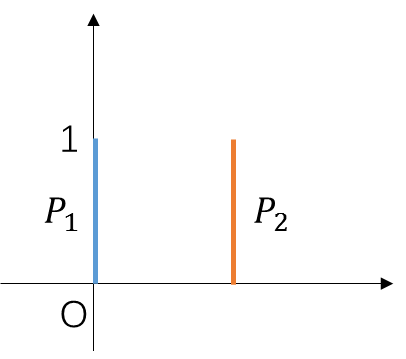
其中



表示以**P**r, **P**g为边缘分布的所有联合分布组成的集合。

我们用一个简单的例子来看一下这四种距离/散度是怎么计算的。

考虑下图的两个均匀分布：



二维平面上，P1是沿着y轴的[0,1]区间上的均匀分布，P2是沿着x=θ，在y轴的[0,1]区间上的均匀分布。简而言之，你可以把P1和P2看成是两条平行的线段。容易计算，



当θ🡪0时，W🡪0，然而TV距离、KL散度、JS散度都不收敛。也就是说，地动距离对某些情况下要更合理一些。更严谨的结论由下面的定理给出。

PS: 为了统一编号，后续的定理编号与原文[2]的编号不一样，两种编号相差3...

**Theorem 4**: 设**P**r是定义在X上的一个固定分布，Z是定义在**Z**空间上的随机变量。再设



是一个函数，记为



记g的分布为**P**θ，

则

1. 若g关于θ连续，则W(**P**r, **P**θ)也连续。
2. 若g满足局部Lipschitz条件，局部Lipschitz常数为L(θ, z)，且



则W(**P**r, **P**θ)处处连续，且几乎处处可微。

1. 上述两个结论对JS散度和KL散度均不成立。

定理4表明，地动距离与JS散度、KL散度相比，具有更好的性质。

**Corollary 4**: 设g\_θ是任意一个前向传播神经网络，带有参数θ，并且p(z)是关于z的先验概率，满足



g满足局部Lipschitz条件，局部Lipschitz常数为L(θ, z)，且



则W(**P**r, **P**θ)处处连续，且几乎处处可微。

推论4表明将地动距离作为神经网络的目标函数是可行的。

**Theorem 5**: 设**P**是紧致空间X上的一个分布，



是X上的一个分布序列，则考虑n🡪∞时，

1. 下面两个命题等价：

(1). δ(**P**n, **P**) 🡪 0

(2). JS(**P**n, **P**) 🡪 0

1. 下述命题等价：

(1). W(**P**n, **P**)🡪0

(2). **P**n依分布收敛于**P**

1. KL(**P**n || **P**) 🡪 0或者KL(**P** || **P**n) 🡪 0能导出结论1.
2. 结论1能导出结论2.

定理5表明，如果分布的支撑集在低维流形上，KL散度、JS散度和TV距离并不是好的loss，而地动（EM）距离则很合适。这启发我们可以用地动距离来设计loss以替换原来GAN采用的KL散度。

**WGAN**

采用Wasserstein距离作为loss的GAN称为Wasserstein GAN，一般简写为WGAN。直接考虑Wasserstein距离需要算inf，计算是很困难的。考虑它的Kantorovich-Rubinstein对偶形式



也就是说，Wasserstein距离实际上需要考虑所有的1-Lipschitz函数。如果我们考虑的是K-Lipschitz函数，则Wasserstein距离变为原来的K倍。一般来说，我们可以假定GAN的discriminator（D）是K-Lipschitz函数，即优化D实际上是在某个K-Lipschitz函数集



上寻找合适的函数（其中w为f的参数）。

进一步地，我们来比较WGAN的D的目标函数（cost/loss）和原来的GAN的目标函数的差别：



可以看到，如果把GAN的目标函数的log去掉，则两者只相差一个常数，也就是说，WGAN在训练的时候与GAN几乎一样，除了loss计算的时候不取对数！Loss function中的对数函数导致了GAN训练的不稳定！

**Theorem 6**: 设**P**r是任一分布，**P**θ是



的分布，其中Z的先验概率密度函数为p(z)，g满足局部Lipschitz条件，局部Lipschitz常数为L(θ, z)，且



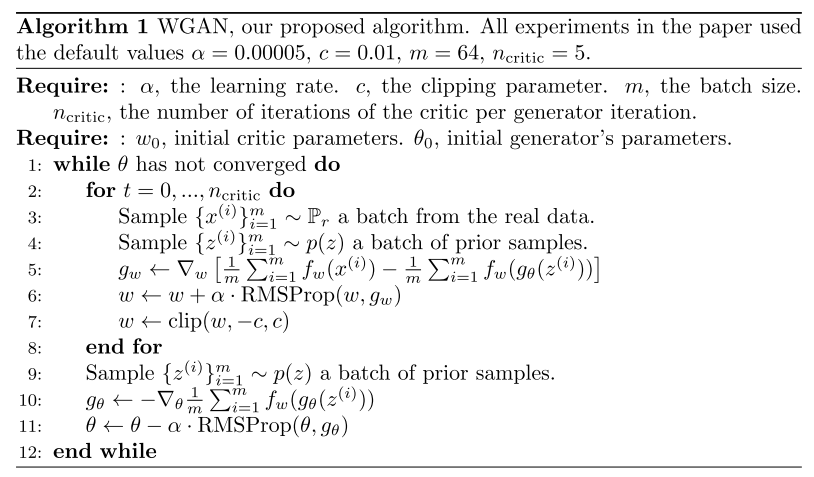
则，若下述结论中的两个式子是良定义的，下述问题



存在解f: X 🡪 R，并且，



定理6证明了若D和G的学习能力足够强的话（因此目标函数能够被最大化），WGAN是有解的。WGAN的算法流程如下：

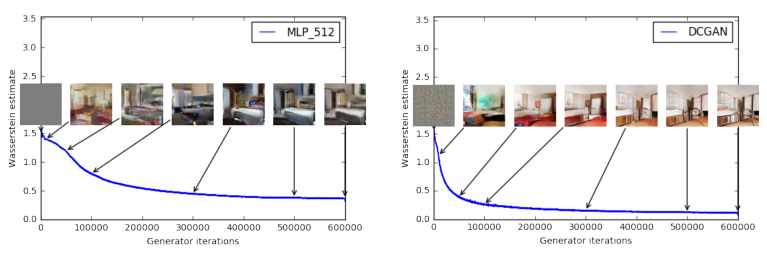


**WGAN实验**

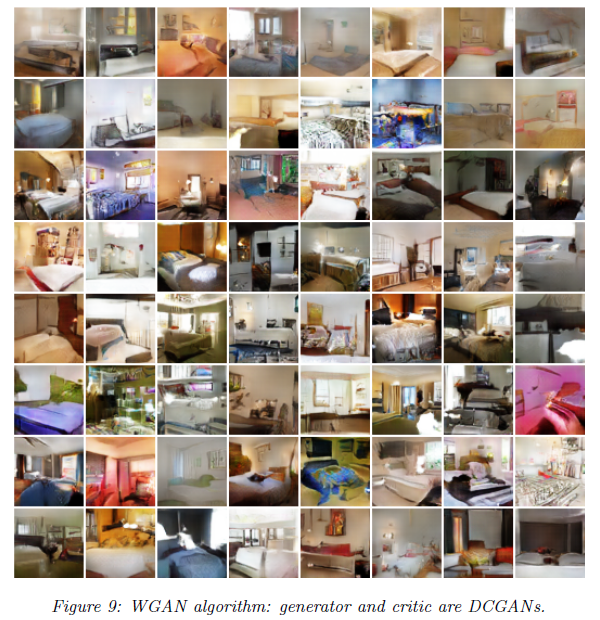
作者发现，如果WGAN训练**采用SGD或者RMSProp算法**，则收敛效果很好。一般**不采用基于momentum的算法，如Adam算法**，实现观察发现这类优化算法会导致训练变得不稳定。而我们知道，DCGAN采用Adam算法进行优化效果会比较好。这是WGAN与GAN训练方法的差别。

此外，**WGAN当前的loss（Wasserstein距离）能够用于指示训练的效果**，即G产生的图像质量，Wasserstein距离越小，G产生的图像质量就越高。先前的GAN由于训练不稳定，我们很难通过loss去判断G产生的质量（先前的GAN的loss大小并不能表明产生图像质量的高低）。这个发现对于训练GAN有很大的帮助。

此外，WGAN如果采用DCGAN架构去训练，产生的图像质量效果与DCGAN没有明显差异；并且，**即使generator采用MLP（多层感知机），仍然能够产生质量不错的图像**。实验结果如下图所示（图中的曲线不会跟GAN一样产生强烈震荡了！）。



此外，作者指出，**WGAN的实验中并没有发现mode collapse**！







**WGAN代码及材料**

Reddit讨论区传送门：

<https://www.reddit.com/r/MachineLearning/comments/5qxoaz/r_170107875_wasserstein_gan/?from=groupmessage>

推荐一篇用更通俗易懂的语言介绍WGAN的文章：

<https://zhuanlan.zhihu.com/p/25071913>

WGAN源码，作者提供，Torch版本：

<https://github.com/martinarjovsky/WassersteinGAN>

Tensorflow版本：<https://github.com/Zardinality/WGAN-tensorflow>

Keras版本：<https://github.com/tdeboissiere/DeepLearningImplementations/tree/master/WassersteinGAN>

**参考文献**

1. Arjovsky, M., & Bottou, L. eon. (2017). Towards Principled Methods for Training Generative Adversarial Networks.
2. Arjovsky, M., Soumith, C., & Bottou, L. eon. (n.d.). Wasserstein GAN.

**下期预告**

看得头晕晕？说实话，定理的证明我也还没有全部弄懂...下期我们来聊聊轻松的话题，主题是GAN的应用，先前没有推送的text to image会在下期中一并推送。当然，GAN的应用不止于此，详情敬请期待下期推送。下期的参考文献我还没整理出来，求不打脸。