瓦瑟斯坦GAN

Martin Arjovsky[1], Soumith Chintala[2], and Léon Bottou[1,2]

[1]Courant Institute of Mathematical Sciences
[2]Facebook AI Research

1 简介

本文研究的是无监督学习问题。主要是，学习概率是什么意思分配?经典的答案是学习概率密度。这通常是通过定义密度的参数族来实现的$(P_\theta)_{\theta\in\mathbb{R}^d}$和找到一个最大的可能性在我们的数据: 如果我们有真实的数据例子$\{x^{(i)}\}_{i=1}^m$，我们就能解决这个问题

$$\max_{\theta\in\mathbb{R}^d}\frac{1}{m}\sum_{i=1}^{m}\log P_\theta(x^{(i)})$$

如果真实数据分布$\mathbb{P}_r$允许密度，并且$\mathbb{P}_\theta$是参数化密度的分布P_θ，那么，渐近地，这等于最小化Kullback-Leibler散度$KL(\mathbb{P}_r\|\mathbb{P}_\theta)$。

为了使这一点有意义，我们需要模型密度P_θ存在。在我们所处的相当普遍的情况下，情况并非如此处理由低维流形支持的分布。那么模型流形和真实分布的支持就不太可能了有一个不可忽略的交集(见[1])，这意味着KL距离没有定义(或者只是无限的)。

典型的补救措施是在模型中加入噪声项分布。这就是为什么中描述的所有生成模型经典的机器学习文献中包含了噪声组件。在最简单的情况下，假设高斯噪声相对较高的带宽，以覆盖所有的例子。是的例如，众所周知，在图像生成模型中，这种噪声降低了样品的质量，使它们模糊的。例如，我们可以在最近的论文[23]中看到即将噪声的最优标准差添加到模型中当最大似然值在生成图像的每个像素的0.1左右时，当像素已经归一化到$[0,1]$范围内时。这是噪音非常大，以至于当报纸报道样本时在他们的模型中，他们没有添加噪声项来报告可能性数。换句话说，附加的噪声项显然是不正确的问题，但需要使最大似然方法有效。

与其估计可能不存在的$\mathbb{P}_r$的密度，我们可以定义一个随机变量Z与固定分布$p(z)$和通过参数函数$g_\theta:\mathcal{Z}\to\mathcal{X}$(通常是某种类型的神经网络)直接传递它生成遵循特定分布$\mathbb{P}_\theta$的示例。By 改变θ，我们可以改变这个分布，使它接近真实数据分布$\mathbb{P}_r$。这在两个方面很有用。首先总之，与密度不同，这种方法可以表示分布局限于低维流形的。第二，轻松的能力生成样本通常比知道数值更有用的密度(例如图像的超分辨率或语义) 分割时考虑条件分布给定输入图像的输出图像)。一般来说，它是计算性的在给定任意高维的情况下很难生成样本密度[16]。

变分自编码器(VAEs) [9]和生成对抗网络(GANs) [4]就是众所周知的例子这种方法。因为VAEs关注的是这些例子都有标准模型的局限性需要修改额外的噪音条

款。gan提供的更多在目标函数的定义上的灵活性，包括Jensen-Shannon [4]和所有f -散度[17]以及一些奇异的组合[6]。另一方面，训练gan是众所周知的是脆弱和不稳定的，理论上的原因在[1]。

在本文中，我们将注意力集中在各种测量方法上模型分布和实际分布有多接近同样地，用不同的方法来定义距离或散度$\rho(\mathbb{P}_\theta, \mathbb{P}_r)$。这两者之间最根本的区别距离是它们对序列收敛性的影响概率分布。分布序列$(\mathbb{P}_t)_{t\in\mathbb{N}}$当且仅当存在分布时收敛$\mathbb{P}_\infty$使得$\rho(\mathbb{P}_t, \mathbb{P}_\infty)$趋于0，这取决于距离$\rho$是如何定义的。非正式地说，距离$\rho$会导致较弱的拓扑结构一个分布序列更容易收敛。[1]章节2 阐明了流行概率距离在这方面的差异。

为了优化参数θ，这当然是可取的定义我们的模型分布$\mathbb{P}_\theta$的方式使映射$\theta \mapsto \mathbb{P}_\theta$连续。连续性是指当参数序列$\theta_t$收敛时对于$\theta$，分布$\mathbb{P}_{\theta_t}$也收敛到$\mathbb{P}_\theta$。然而，重要的是要记住分布收敛的概念$\mathbb{P}_{\theta_t}$取决于在计算分布之间距离的过程中。这种距离越弱，就越容易定义一个从θ -space到$\mathbb{P}_\theta$ -空间，因为它更容易分布收敛。我们关心映射的主要原因是$\theta \mapsto \mathbb{P}_\theta$ To be continuous是连续的。如果ρ是我们的概念两个分布之间的距离，我们要有一个损失函数$\theta \mapsto \rho(\mathbb{P}_\theta, \mathbb{P}_r)$是连续的，这等价于映射$\theta \mapsto \mathbb{P}_\theta$使用分布之间的距离时要连续$\rho$。

[1]更多信息没错，ρ诱导的拓扑比诱导的要弱当ρ下的收敛序列集合为a时，通过ρ' ρ'下的超集。

本文的贡献有:

- 在2节中，我们提供了一个全面的理论分析了地动器(EM)距离的变化规律流行的概率距离和散度用于学习分布的背景。
- 在3节中，我们定义了一种称为Wasserstein-GAN的GAN形式最小化合理有效的电磁距离近似值，并从理论上证明了相应的优化问题是声音。
- 在4节中，我们实证地证明了wgan 解决了gan的主要训练问题。特别是培训wgan不需要在训练中保持谨慎的平衡鉴别器和生成器一样，不需要小心网络架构的设计。模式下降gan中典型的现象也大大减少了。一wgan最引人注目的实际好处之一是能够通过训练，连续估计EM距离鉴别器到最优性。绘制这些学习曲线则不然仅对调试和超参数搜索有用，但也与观察到的样品质量显著相关。

2 不同的距离

现在我们介绍我们的符号。设$\mathcal{X}$是紧度量集(例如图像的空间$[0,1]^d$)，并让Σ表示$\mathcal{X}$的所有Borel子集的集合。让$\text{Prob}(\mathcal{X})$ 表示在$\mathcal{X}$上定义的概率测度的空间。现在我们可以定义初等距离和散度了在两个发行版之间$\mathbb{P}_r, \mathbb{P}_g \in \text{Prob}(\mathcal{X})$:

- 总变差(TV)距离
$$\delta(\mathbb{P}_r, \mathbb{P}_g) = \sup_{A \in \Sigma} |\mathbb{P}_r(A) - \mathbb{P}_g(A)| \ .$$
- *Kullback-Leibler*散度
$$KL(\mathbb{P}_r \| \mathbb{P}_g) = \int \log\left(\frac{P_r(x)}{P_g(x)}\right) P_r(x) d\mu(x) \ ,$$
假设$\mathbb{P}_r$和$\mathbb{P}_g$是绝对连续的，因此承认密度，相对于相同的测量μ定义在$\mathcal{X}$上。[2] KL散度是出了名的不对称，在那里可能是无限的是这样的点$P_g(x) = 0$和$P_r(x) > 0$。
- *Jensen-Shannon* (JS)散度
$$JS(\mathbb{P}_r, \mathbb{P}_g) = KL(\mathbb{P}_r \| \mathbb{P}_m) + KL(\mathbb{P}_g \| \mathbb{P}_m) \ ,$$
其中$\mathbb{P}_m$是混合物$(\mathbb{P}_r + \mathbb{P}_g)/2$。这种分歧是对称的，并且总是有定义的，因为我们可以选择$\mu = \mathbb{P}_m$。
- 土动器(EM)距离或Wasserstein-1
$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma \in \Pi(\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(x,y)\sim\gamma}\big[\, \|x - y\| \,\big] \ , \tag{1}$$
其中$\Pi(\mathbb{P}_r, \mathbb{P}_g)$表示所有联合分布的集合$\gamma(x, y)$，其边际分别为$\mathbb{P}_r$和$\mathbb{P}_g$。直观地，$\gamma(x, y)$表示必须有多少“质量” 为了转换发行版，从$x$ 传输到y $\mathbb{P}_r$到分布$\mathbb{P}_g$中。然后是电磁距离是最优运输方案的“成本”。

[2]回忆一下概率分布$\mathbb{P}_r \in \text{Prob}(\mathcal{X})$承认密度$p_r(x)$ with 对于μ，即$\forall A \in \Sigma$，$\mathbb{P}_r(A) = \int_A P_r(x) d\mu(x)$，当且仅当它是绝对连续的对于$\mu$，也就是$\forall A \in \Sigma$，$\mu(A) = 0 \Rightarrow \mathbb{P}_r(A) = 0$。

下面的例子说明了看似简单的序列概率分布在电磁距离下收敛但不收敛收敛于上述定义的其他距离和散度下。

Example 1 (Learning parallel lines)**.** Let $Z \sim U[0,1]$ the uniform distribution on the unit interval. Let $\mathbb{P}_0$ be the distribution of $(0, Z) \in \mathbb{R}^2$ (a 0 on the x-axis and the random variable Z on the y-axis), uniform on a straight vertical line passing through the origin. Now let $g_\theta(z) = (\theta, z)$ with θ a single real parameter. It is easy to see that in this case,

- $W(\mathbb{P}_0, \mathbb{P}_\theta) = |\theta|$,
- $JS(\mathbb{P}_0, \mathbb{P}_\theta) = \begin{cases} \log 2 & \text{if } \theta \neq 0\ , \\ 0 & \text{if } \theta = 0\ , \end{cases}$
- $KL(\mathbb{P}_\theta \| \mathbb{P}_0) = KL(\mathbb{P}_0 \| \mathbb{P}_\theta) = \begin{cases} +\infty & \text{if } \theta \neq 0\ , \\ 0 & \text{if } \theta = 0\ , \end{cases}$
- and $\delta(\mathbb{P}_0, \mathbb{P}_\theta) = \begin{cases} 1 & \text{if } \theta \neq 0\ , \\ 0 & \text{if } \theta = 0\ . \end{cases}$

When $\theta_t \to 0$, the sequence $(\mathbb{P}_{\theta_t})_{t \in \mathbb{N}}$ converges to $\mathbb{P}_0$ under the EM distance, but does not converge at all under either the JS, KL, reverse KL, or TV divergences. Figure 1 illustrates this for the case of the EM and JS distances.

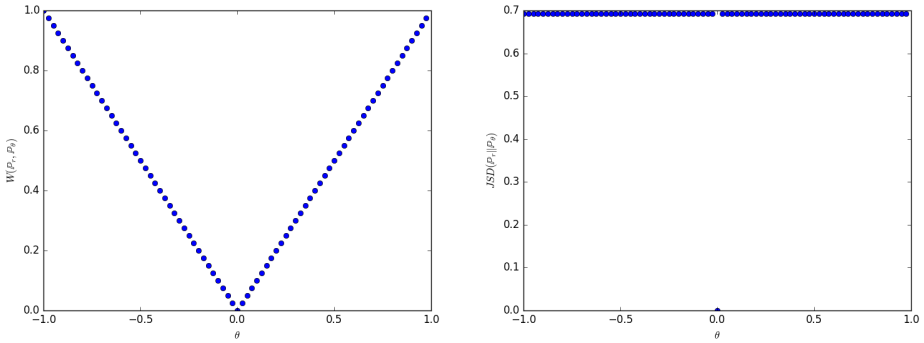

Figure 1: 这些图表显示$\rho(\mathbb{P}_\theta, \mathbb{P}_0)$是$\theta$的函数当$\rho$为*EM*距离*(*左图*)*或*JS*散度*(*右图*)*。*EM*图是连续的，并在任何地方提供可用的梯度。*JS*图不是连续的，也没有提供可用的梯度。

例子1给了我们一个可以学习a的例子在低维流形上的概率分布EM距离的梯度下降。这不能用其他距离和散度，因为最终的损失函数是甚至不是连续的。虽然这个简单的例子具有具有不相交支持的分布，当的集合中包含一个非空的交叉点测量零。这恰好是两个低维空间的情况歧管在一般位置相交[1]。

由于Wasserstein距离远小于JS距离[3]，我们现在可以问$W(\mathbb{P}_r, \mathbb{P}_\theta)$是否在$\theta$上的连续损失函数在温和的假设下。这个，还有更多，是没错，正如我们现在陈述和证明的那样。

Theorem 1. *Let $\mathbb{P}_r$ be a fixed distribution over $\mathcal{X}$. Let Z be a random variable (e.g Gaussian) over another space $\mathcal{Z}$. Let $g : \mathcal{Z} \times \mathbb{R}^d \to \mathcal{X}$ be a function, that will be denoted $g_\theta(z)$ with z the first coordinate and θ the second. Let $\mathbb{P}_\theta$ denote the distribution of $g_\theta(Z)$. Then,*

1. *If g is continuous in θ, so is $W(\mathbb{P}_r, \mathbb{P}_\theta)$.*
2. *If g is locally Lipschitz and satisfies regularity assumption 1, then $W(\mathbb{P}_r, \mathbb{P}_\theta)$ is continuous everywhere, and differentiable almost everywhere.*
3. *Statements 1-2 are false for the Jensen-Shannon divergence $JS(\mathbb{P}_r, \mathbb{P}_\theta)$ and all the KLs.*

Proof. 参见附录C □

下面的推论告诉我们通过最小化来学习EM距离对于神经网络来说是有意义的(至少在理论上)。

Corollary 1. *Let g_θ be any feedforward neural network[4] parameterized by θ, and $p(z)$ a prior over z such that $\mathbb{E}_{z\sim p(z)}[\|z\|] < \infty$ (e.g. Gaussian, uniform, etc.). Then assumption 1 is satisfied and therefore $W(\mathbb{P}_r, \mathbb{P}_\theta)$ is continuous everywhere and differentiable almost everywhere.*

Proof. 参见附录C □

所有这些都表明，新兴市场要明智得多我们问题的成本函数至少比Jensen-Shannon函数要好散度。的相对强度由这些距离和散度引起的拓扑，其中KL最强; 其次是JS和TV, EM最弱。

Theorem 2. *Let $\mathbb{P}$ be a distribution on a compact space $\mathcal{X}$ and $(\mathbb{P}_n)_{n\in\mathbb{N}}$ be a sequence of distributions on $\mathcal{X}$. Then, considering all limits as $n \to \infty$,*

1. *The following statements are equivalent*
   - *$\delta(\mathbb{P}_n, \mathbb{P}) \to 0$ with δ the total variation distance.*
   - *$JS(\mathbb{P}_n, \mathbb{P}) \to 0$ with JS the Jensen-Shannon divergence.*
2. *The following statements are equivalent*
   - *$W(\mathbb{P}_n, \mathbb{P}) \to 0$.*

[3]为什么会发生这种情况，事实上我们是怎么想到沃瑟斯坦是什么的我们真正应该优化的是显示在附录中Appendix A。我们强烈鼓励有兴趣的人不惧怕数学的读者要通读它。

[4]By a feedforward neural network we mean a function composed by affine transformations and pointwise nonlinearities which are smooth Lipschitz functions (such as the sigmoid, tanh, elu, softplus, etc). **Note**: the statement is also true for rectifier nonlinearities but the proof is more technical (even though very similar) so we omit it.

- $\mathbb{P}_n \xrightarrow{\mathcal{D}} \mathbb{P}$ *where* $\xrightarrow{\mathcal{D}}$ *represents convergence in distribution for random variables.*

3. $KL(\mathbb{P}_n \| \mathbb{P}) \to 0$ *or* $KL(\mathbb{P} \| \mathbb{P}_n) \to 0$ *imply the statements in (1).*

4. *The statements in (1) imply the statements in (2).*

Proof. 参见附录C □

这凸显了一个事实KL, JS和TV距离是不明智的学习分布时的代价函数由低维流形支撑。然而，电磁距离是合理的在这种情况下。这显然将我们引向下一节我们在哪里引入一个实用的近似优化电磁距离。

3 瓦瑟斯坦GAN

定理2再次指出了这样一个事实$W(\mathbb{P}_r, \mathbb{P}_\theta)$可能有更好的属性优化时比$JS(\mathbb{P}_r, \mathbb{P}_\theta)$。然而，在(1)是非常棘手的。另一方面，坎托罗维奇-鲁宾斯坦二象性[22] 告诉我们

$$W(\mathbb{P}_r, \mathbb{P}_\theta) = \sup_{\|f\|_L \leq 1} \mathbb{E}_{x \sim \mathbb{P}_r}[f(x)] - \mathbb{E}_{x \sim \mathbb{P}_\theta}[f(x)] \tag{2}$$

在所有的1-Lipschitz函数上极值在哪里$f : \mathcal{X} \to \mathbb{R}$。注意，如果我们替换$\|f\|_L \leq 1$对于$\|f\|_L \leq K$(考虑$K$ -Lipschitz对于某些常数K)，然后我们最终得到$K \cdot W(\mathbb{P}_r, \mathbb{P}_g)$。因此，如果我们有参数化的函数族$\{f_w\}_{w \in \mathcal{W}}$ 都是K -Lipschitz对于一些K，我们可以考虑解决问题

$$\max_{w \in \mathcal{W}} \mathbb{E}_{x \sim \mathbb{P}_r}[f_w(x)] - \mathbb{E}_{z \sim p(z)}[f_w(g_\theta(z)] \tag{3}$$

如果达到(2)的至高点对于一些$w \in \mathcal{W}$(一个相当强的假设类似于证明an的一致性时的假设估计器)，这个过程将产生一个计算$W(\mathbb{P}_r, \mathbb{P}_\theta)$到一个乘法常数。此外，我们可以考虑对$W(\mathbb{P}_r, \mathbb{P}_\theta)$求导(同样，直到一个常数) 通过方程(2) via进行反向支撑估算$\mathbb{E}_{z \sim p(z)}[\nabla_\theta f_w(g_\theta(z))]$。虽然这些都是直觉，我们现在证明这个过程在最优假设下是原则性的。

Theorem 3. *Let $\mathbb{P}_r$ be any distribution. Let $\mathbb{P}_\theta$ be the distribution of $g_\theta(Z)$ with Z a random variable with density p and g_θ a function satisfying assumption 1. Then, there is a solution $f : \mathcal{X} \to \mathbb{R}$ to the problem*

$$\max_{\|f\|_L \leq 1} \mathbb{E}_{x \sim \mathbb{P}_r}[f(x)] - \mathbb{E}_{x \sim \mathbb{P}_\theta}[f(x)]$$

and we have

$$\nabla_\theta W(\mathbb{P}_r, \mathbb{P}_\theta) = -\mathbb{E}_{z \sim p(z)}[\nabla_\theta f(g_\theta(z))]$$

when both terms are well-defined.

Proof. 参见附录Appendix C □

Algorithm 1 我们提出的算法WGAN。所有实验都在论文中使用默认值$\alpha = 0.00005$、$c = 0.01$、$m = 64$、$n_{\text{critic}} = 5$。

Require: : α, the learning rate. c, the clipping parameter. m, the batch size. n_{critic}, the number of iterations of the critic per generator iteration.

Require: : w_0, initial critic parameters. θ_0, initial generator's parameters.

while θ has not converged **do**
  for $t = 0, ..., n_{\text{critic}}$ **do**
    Sample $\{x^{(i)}\}_{i=1}^{m} \sim \mathbb{P}_r$ a batch from the real data.
    Sample $\{z^{(i)}\}_{i=1}^{m} \sim p(z)$ a batch of prior samples.
    $g_w \leftarrow \nabla_w \left[\frac{1}{m}\sum_{i=1}^{m} f_w(x^{(i)}) - \frac{1}{m}\sum_{i=1}^{m} f_w(g_\theta(z^{(i)}))\right]$
    $w \leftarrow w + \alpha \cdot \text{RMSProp}(w, g_w)$
    $w \leftarrow \text{clip}(w, -c, c)$
  end for
  Sample $\{z^{(i)}\}_{i=1}^{m} \sim p(z)$ a batch of prior samples.
  $g_\theta \leftarrow -\nabla_\theta \frac{1}{m}\sum_{i=1}^{m} f_w(g_\theta(z^{(i)}))$
  $\theta \leftarrow \theta - \alpha \cdot \text{RMSProp}(\theta, g_\theta)$
end while

现在的问题是找到函数f 求解方程(2)中的最大化问题。大致近似我们可以训练一个神经网络用权重参数化w位于紧空间中$\mathcal{W}$然后反向通过$\mathbb{E}_{z\sim p(z)}[\nabla_\theta f_w(g_\theta(z))]$，作为我们用一个典型的GAN来做。注意这个事实$\mathcal{W}$是紧凑的意味着所有的功能$f_w$将是$K$ -Lipschitz对于一些K，这取决于情况在$\mathcal{W}$上，而不是个人权重近似(2)到一个不相关的比例因子以及"批评家"的能力f_w。为了有参数w 在一个紧凑的空间里，我们能做的一件简单的事就是夹紧权重到一个固定的盒子(比如$\mathcal{W} = [-0.01, 0.01]^l$)之后梯度更新。沃瑟斯坦生成对抗网络(WGAN)过程在算法1中描述。

权重裁剪显然是一种强制Lipschitz约束的糟糕方式。如果裁剪参数较大，则需要较长的时间使任何重量达到极限，从而使它变得更加困难训练批评家直到达到最佳状态。如果剪辑很小，则为当层数很大时，很容易导致梯度消失，或者不使用批处理归一化(例如在rnn中)。我们进行了实验使用简单的变体(例如将权重投射到球体上) 差别不大，我们坚持使用重量裁剪，因为它很简单而且已经有很好的表现了。然而，我们离开了这个话题进一步在神经网络设置中强制Lipschitz约束调查，我们积极鼓励感兴趣的研究人员改进就这个方法。

EM距离连续可微的事实。意味着我们可以(也应该)训练批评家直到最优状态。这个论点很简单，我们对批评家的训练越多，他们的梯度就越可靠我们得到的沃瑟斯坦，它实际上是有用的沃瑟斯坦几乎在任何地方都是可微的。对于JS，随着鉴别器变得更好，梯度变得更可靠，但真正的梯度是0，因为JS是本地的饱和，然后渐变消失，从本文的Figure 1可以看出和[1]的定理2.4。在Figure 2 我们展示了这个概念的证明，我们训练的地方一个GAN鉴别器和一个WGAN最优性评价器。鉴别器很快就学会了区分假与真，并如预期提供不可靠的梯度信息。然而，批评家不能饱和，而要趋同到一个线性函数，它在任何地方都有非常清晰的梯度。我们对权重的限制限制了可能性函数在不同部分的增长最多是线性的的空间，迫使最佳评论家有这种行为。

也许更重要的是，我们可以训练批评家直到当我们这样做时，最优性使其不可

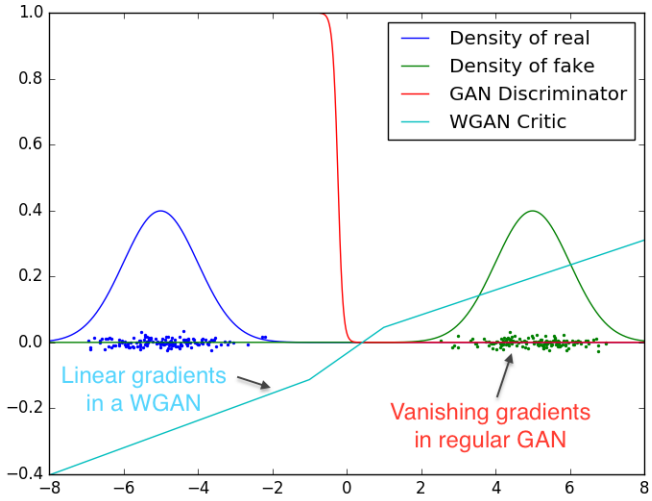

Figure 2: 最佳鉴别器和批评家学习区分两个高斯函数。我们可以看到，极小极大*GAN*的鉴别器饱和并得到结果在消失的梯度中。我们的*WGAN*评论家提供非常干净空间所有部分的渐变。

能崩溃模式。这是由于模态塌缩来自于这样一个事实对于一个固定鉴别器的最优发生器是鉴别器指定的点上的函数的和吗最大值，如[4]和在[11]中突出显示。

在下一节中，我们将展示实际的好处我们的新算法，我们提供了一个深入的比较它的行为和传统gan的行为。

4 实证结果

我们使用Wasserstein-GAN算法进行图像生成实验并证明使用它有显著的实际好处标准gan中使用的配方。

我们声称有两个主要好处:

- 与发电机相关的有意义的损耗指标收敛性和样本质量
- 提高了优化过程的稳定性

4.1 实验步骤

我们进行图像生成实验。学习的目标分布是lsun -卧室数据集[24]——自然图像的集合室内卧室。我们的基线比较是DCGAN [18]，使用标准GAN程序训练的具有卷积结构的GAN 使用$-\log D$技巧[4]。生成的样本是大小为64x64像素的3通道图像。我们使用Algorithm 1中指定的超参数我们所有的实验。

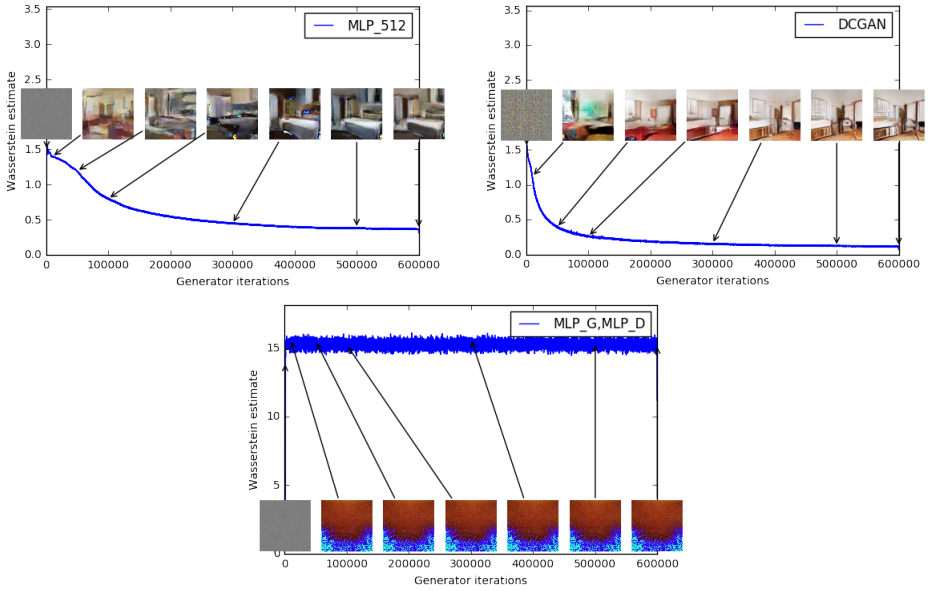

Figure 3: 不同阶段的训练曲线和样本训练。我们可以看到一个明显的相关性误差和更好的样品质量。左上:生成器是一个隐藏了*4*的*MLP* 层，每层*512*个单位。损耗减小随着培训进度和样品质量的不断提高增加。右上:发电机是标准的*DCGAN*。损耗迅速降低，样品质量提高也在上面两幅图中，批评家都是没有乙状结肠的*DCGAN*，因此损失可以进行比较。下半部分:生成器和鉴别器都是*mlp* 具有相当高的学习率*(*所以训练失败了*)*。损失是常数，样本也是常数。培训曲线通过中值过滤器进行可视化处理目的。

4.2 有意义损失度量

因为WGAN算法试图训练评论家f(算法1中的第2- 8行) 在每次生成器更新之前(算法1中的第10行)，此时的损失函数是EM距离的估计值，直至常数与我们约束Lipschitz常数f的方式有关的因素。

我们的第一个实验说明了这个估计是如何很好地相互关联的生成的样本的质量。除了卷积DCGAN架构，我们也做了一些实验生成器或者生成器和评论家都有4层带有512个隐藏单元的ReLU-MLP。

Figure 3 绘制了WGAN估计的演化图三者在WGAN训练期间EM距离的(3) 架构。图表清楚地表明这些曲线是相互关联的生成样本的视觉质量。

据我们所知，这是GAN文献中第一次发现这样的属性，其中氮化镓的损失显示为收敛性。这个属性在做研究时非常有用对抗性网络作为一个不需要盯着生成的样品来找出故障模式并获得信息模特比其他人做得更好。

然而，我们并不声称这是一种新的定量方法评估生成模型。常数缩放取决于评论家的架构的因素意味着它很难比较不同评论家的模型。甚至，实际上，批评家并没有无限的容量让我们很难知道其中的原因接近我们估计的电磁距离。话虽如此，我们已经成功地使用损耗度量来反复验证我们的实验，没有失败，并且我们认为这是训练gan的巨大进步以前没有这样的设施。

相反，Figure 4绘制了GAN估计的演变GAN训练过程中JS距离的。更准确地

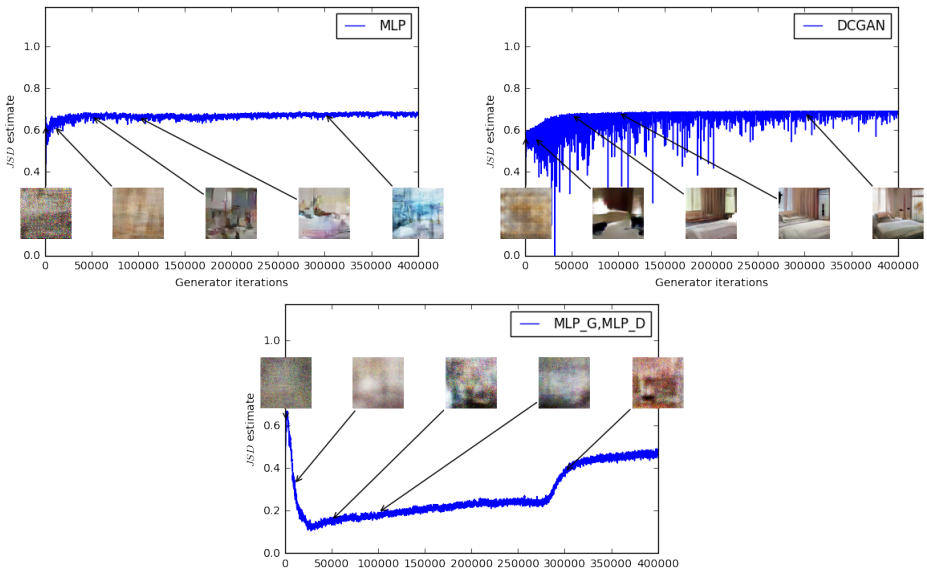

*Figure 4: JS MLP*发生器的估计*(*左上*)* 以及经过标准训练的*DCGAN*发生器*(*右上*) GAN*程序。两者都有一个*DCGAN*鉴别器。两条曲线误差越来越大。*DCGAN*的样本变得更好了但*JS*估计增加或保持不变，指向样品质量与损失之间无显著相关性。底部*: MLP*同时具有生成器和鉴别器。曲线有上有下无论样品质量如何。所有培训曲线通过与*Figure 3*相同的中值过滤器。

说，在GAN训练过程中，鉴别器被训练为最大化

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

也就是$2JS(\mathbb{P}_r, \mathbb{P}_\theta) - 2\log 2$的下界。在图中，我们绘制了数量$\frac{1}{2}L(D, g_\theta) + \log 2$，它是JS距离的下界。

这个数量显然与样品质量关系很差。说明而且JS估计通常保持不变或上升向下走。事实上它通常非常接近$\log 2 \approx 0.69$，这是最高的JS距离取的值。换句话说，就是JS距离饱和，鉴别器具有零损耗，和生成的样本在某些情况下是有意义的(DCGAN生成器，右上方的图) 其他情况下崩溃成一个荒谬的图像[4]。最后一种现象在理论上已被解释[1]并在[11]中突出显示。

当使用$-\log D$技巧[4]时，鉴别器损耗和发生器损耗是不同的。参见附录E中的8为GAN训练报告相同的图，但使用发生器损耗而不是鉴别器损耗。这不会改变结论。

最后，作为负面结果，我们报告WGAN训练变得不稳定有时当一个人使用基于动量的优化器，如Adam [8](与$\beta_1 > 0$) 对评论家来说，还是当一个人使用高学习率时。因为批评家的损失是不稳定的，动量基于方法的表现似乎更差。我们确定了动量是一个潜在的原因，因为随着损失的扩大和样本的恶化，亚当步长和梯度之间的余弦值通常变为负值。The 只有在这种情况下cos是负的不稳定性。因此，我们切换到RMSProp [21]，这是已知的即使在非常非平稳的问题上也能表现得很好[13]。

Figure 5: 用*DCGAN*生成器训练的算法。左*:WGAN*算法。右*:* 标准*GAN*配方。两种算法都能产生高质量的样本。

Figure 6: 算法训练与生成器没有批处理规范化每一层都有固定数量的过滤器*(*而不是像*[18]*那样每次都复制它们*)*。旁白从取出批归一化，参数的数量因此是减少了一个数量级多一点。左*:WGAN*算法。右*:*标准*GAN*配方。我们可以看到标准*GAN*学习失败，而*WGAN*仍然能够学习生产样品。

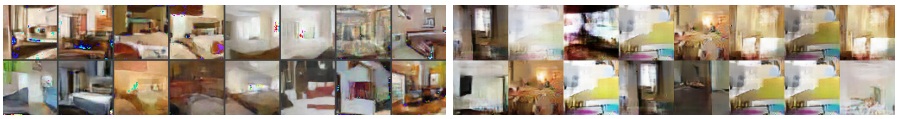

Figure 7: 算法训练的*MLP*生成器有*4*层和具有*ReLU*非线性的*512*个单元。参数的数量与*DCGAN*相似，但它缺乏强烈的归纳偏置用于图像生成。左*:WGAN*算法。右*:*标准*GAN*配方。*WGAN*方法仍然能够产生样品，但质量低于*DCGAN*，并且比标准*GAN*的*MLP*质量更高。注意它的重要性*GAN MLP*中模式崩溃的程度。

4.3 提高稳定性

WGAN的一个好处是，它允许我们训练评论家直到最优。当批评家被训练完成时，它很简单给生成器提供一个损失，我们可以像训练其他神经系统一样训练它网络。这告诉我们，我们不再需要平衡发电机和鉴别器的能力。批评家越好，评价越高质量我们用来训练生成器的梯度。

我们观察到，当一个变量变化时，wgan比gan更具鲁棒性生成器的体系结构选择。我们来说明一下通过在三种生成器架构上运行实验: (1)卷积DCGAN发生器，(2)卷积DCGAN发生器如果没有批处理归一化并且使用恒定数量的过滤器，(3)具有512个隐藏单元的4层ReLU-MLP。已知后两种方法在gan上的表现非常差。我们保留了卷积dgan架构WGAN批评家或GAN鉴别器。

如图5、6和7所示使用WGAN为这三种体系结构生成的示例和GAN算法。我们建议读者参阅附录Appendix F获取生成样本的完整页。样品并不是精心挑选的。

在任何实验中，我们都没有看到WGAN算法模式崩溃的证据。

5 相关工作

有很多关于所谓的积分概率度量(IPMs) [15]。给定$\mathcal{F}$一组来自$\mathcal{X}$的函数对于$\mathbb{R}$，我

们可以定义

$$d_{\mathcal{F}}(\mathbb{P}_r, \mathbb{P}_\theta) = \sup_{f \in \mathcal{F}} \mathbb{E}_{x \sim \mathbb{P}_r}[f(x)] - \mathbb{E}_{x \sim \mathbb{P}_\theta}[f(x)] \tag{4}$$

作为一个积分概率度量与函数类$\mathcal{F}$。很容易证实对于每个$f \in \mathcal{F}$，我们有$-f \in \mathcal{F}$(例如我们将考虑的所有示例)，那么$d_{\mathcal{F}}$是非负的，满足三角形不等式，并且是对称的。因此，$d_{\mathcal{F}}$是在Prob($\mathcal{X}$)上的伪度量。

虽然ipm似乎有相似的公式，我们会看到不同类型的函数可以使用完全不同的指标。

- 通过坎托罗维奇-鲁宾斯坦对偶[22]，我们知道$W(\mathbb{P}_r, \mathbb{P}_\theta) = d_{\mathcal{F}}(\mathbb{P}_r, \mathbb{P}_\theta)$当$\mathcal{F}$是1-Lipschitz集合时函数。此外，如果$\mathcal{F}$是集合$K$ -Lipschitz函数，我们得到$K \cdot W(\mathbb{P}_r, \mathbb{P}_\theta) = d_{\mathcal{F}}(\mathbb{P}_r, \mathbb{P}_\theta)$。

- 当$\mathcal{F}$是所有可测量的集合范围在-1到1(或所有)之间的函数在-1和1之间的连续函数，我们检索$d_{\mathcal{F}}(\mathbb{P}_r, \mathbb{P}_\theta) = \delta(\mathbb{P}_r, \mathbb{P}_\theta)$总变异距离[15]。这已经告诉我们从1-李普希茨开始到1-Bounded函数彻底改变了空间的拓扑结构和规律性$d_{\mathcal{F}}(\mathbb{P}_r, \mathbb{P}_\theta)$的损失函数(如定理1) 2)。

- 基于能量的gan (ebgan) [25] 能想到的作为总变分的生成方法距离。这种联系是声明和在附录Appendix D中有深入的证明。连接的核心是鉴别器会起到f最大化方程的作用吗(4)而它唯一的限制是介于0和m之间某个常数m。这将是相同的行为被限制在-1之间和1到一个常数的比例因子无关优化。因此，当鉴别器逼近发电机成本的最优性是否近似于总变异距离$\delta(\mathbb{P}_r, \mathbb{P}_\theta)$。

  由于总变化距离显示与JS相同的规律，可以看出ebgan也将面临同样的问题关于不能的经典gan 训练鉴别器直到最优从而限制了自己的不完美梯度。

- 最大平均差异(MMD) [5]为积分概率度量的具体情况当$\mathcal{F} = \{f \in \mathcal{H} : \|f\|_\infty \leq 1\}$ for $\mathcal{H}$一些再现核希尔伯特空间(RKHS) 与给定内核关联$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$。正如[5]所证明的那样，我们知道MMD是正确的度量而不仅仅是伪度量，当内核是通用的。在具体的情况下$\mathcal{H} = L^2(\mathcal{X}, m)$表示$m$归一化的勒贝格测量$\mathcal{X}$，我们知道$\{f \in C_b(\mathcal{X}), \|f\|_\infty \leq 1\}$ 将包含在$\mathcal{F}$中，因此$d_{\mathcal{F}}(\mathbb{P}_r, \mathbb{P}_\theta) \leq \delta(\mathbb{P}_r, \mathbb{P}_\theta)$那么MMD距离的规律性呢作为损失函数，它至少会和总的损失函数一样糟糕变异。然而，这是一个非常极端的例子，因为我们会需要一个非常强大的内核来近似整个L^2。然而，即使是高斯核也能检测到微小的噪声模式最近，[20]证明了这一点。这指向了事实上，特别是对于低带宽的内核，距离可能接近饱和状态，类似于total JS的变体。这显然不需要每个内核的情况，并找出如何以及哪些不同的mmd 更接近Wasserstein还是总变异距离是一个有趣的问题研究课题。

  MMD的伟大之处在于，通过内核技巧，没有必要这样做训练一个单独的网络来最大化球的方程(4) RKHS的。然而，这样做的缺点是无法评估MMD距离计算成本是否会随着样本数量呈二次增长用于估算(4)中的期望值。最后一点使得MMD具有有限的可伸缩性，并且有时不适用于许多现实生活中的应用都离不开它。有一些估计线性计算成本为MMD [5] 在很多情况下使MMD非常有用，但它们也具有更差的样本复杂性。

- 生成矩匹配网络(GMMNs) [10, 2] 是MMD的生成对应物。通过反向支撑方程(4)的核化公式，他们直接优化$d_{MMD}(\mathbb{P}_r, \mathbb{P}_\theta)$(当$\mathcal{F}$为时的IPM) 如前一项)。如前所述，这有其优点不需要单独的网络来近似最大化方程(4)。然而，转基因转基因食品的适用性有限。其失败的部分解释是作为函数的二次成本低带宽核的样本数和消失梯度。此外，有可能有些人实际使用的核函数不适合捕获非常复杂的数据高维样本空间(如自然图像)中的距离。这是合理的，因为[19] 表明，对于典型的高斯MMD测试是可靠的(因为它的功率作为一个接近1)的统计检验，我们需要的是样本随维数线性增长。自从MMD的计算成本随数量呈二次增长用于估计方程(4)的批次样本中，这使得拥有一个可靠的估算器的成本增加与维数成二次增长，这使得它非常不适用于高维问题。的确，对于对于像64 x 64这样标准的图像，我们需要minibatch 大小至少4096(不考虑常数) 在[19]的范围内也就是这个数(实质上更大)和每次迭代的总成本4096^2，比a大5个数量级GAN迭代时使用的标准批大小为64。

  话虽如此，这些数字对MMD可能有点不公平，在这个意义上我们比较的是经验样本的复杂度GANs的理论样本复杂度为MMDs，趋于更糟。然而，在最初的GMMN论文[10]中，他们实际上使用了1000的小批量，比标准的大得多32或64(即使这会产生二次计算成本)。而估计的计算代价有线性函数的数目的样本存在[5]，它们的样本复杂度更差，而且是最好的据我们所知，它们还没有被应用到生成环境中比如转基因玉米。

在另一个伟大的研究领域，最近的工作[14] 探索了沃瑟斯坦距离在学习中的应用对于离散空间的受限玻尔兹曼机。动机乍一看可能会有很大的不同，因为歧管设置是受限制的对于连续空间和有限离散空间的弱和强拓扑结构(分别是W和JS)一致。然而，在最后，我们的动机有更多的共同点。我们俩想要以一种利用几何图形的方式来比较分布底层空间，沃瑟斯坦允许我们这样做。

最后，[3]的工作展示了新的计算算法不同分布之间的沃瑟斯坦距离。我们相信这个方向是非常重要的，并且可能会导致评估生成模型的新方法。

6 结论

我们引入了一种算法，我们称之为WGAN，一种替代方案到传统的GAN训练。在这个新模型中，我们证明了我们可以提高学习的稳定性，解决模式崩溃等问题，并提供了对调试和超参数化有用的有意义的学习曲线搜索。进一步，我们证明了相应的优化问题是健全的，并提供了广泛的理论工作突出与分布之间其他距离的深度连接。

致谢

我们要感谢Mohamed Ishmael Belghazi，艾米丽·丹顿，伊恩·古德费罗，Ishaan Gulrajani，亚历克斯·兰姆，大卫·洛佩斯，埃里克·马丁，马克西姆·奥克布，Aditya Ramesh，罗南·里奥谢，乌里·沙利特，Pablo Sprechmann，亚瑟·斯拉姆，王若涵，有帮助的意见和建议。

References

[1] Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. In *International Conference on Learning Representations*, 2017. Under review.

[2] Gintare Karolina Dziugaite, Daniel M. Roy, and Zoubin Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. *CoRR*, abs/1505.03906, 2015.

[3] Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3440–3448. Curran Associates, Inc., 2016.

[4] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

[5] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773, 2012.

[6] Ferenc Huszar. How (not) to train your generative model: Scheduled sampling, likelihood, adversary? *CoRR*, abs/1511.05101, 2015.

[7] Shizuo Kakutani. Concrete representation of abstract (m)-spaces (a characterization of the space of continuous functions). *Annals of Mathematics*, 42(4):994–1024, 1941.

[8] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[9] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013.

[10] Yujia Li, Kevin Swersky, and Rich Zemel. Generative moment matching networks. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 1718–1727. JMLR Workshop and Conference Proceedings, 2015.

[11] Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled generative adversarial networks. *Corr*, abs/1611.02163, 2016.

[12] Paul Milgrom and Ilya Segal. Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2):583–601, 2002.

[13] Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 1928–1937, 2016.

[14] Grégoire Montavon, Klaus-Robert Müller, and Marco Cuturi. Wasserstein training of restricted boltzmann machines. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3718–3726. Curran Associates, Inc., 2016.

[15] Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

[16] Radford M. Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, April 2001.

[17] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. pages 271–279, 2016.

[18] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015.

[19] Aaditya Ramdas, Sashank J. Reddi, Barnabas Poczos, Aarti Singh, and Larry Wasserman. On the high-dimensional power of linear-time kernel two-sample testing under mean-difference alternatives. *Corr*, abs/1411.6314, 2014.

[20] Dougal J Sutherland, Hsiao-Yu Tung, Heiko Strathmann, Soumyajit De, Aaditya Ramdas, Alex Smola, and Arthur Gretton. Generative models and model criticism via optimized maximum mean discrepancy. In *International Conference on Learning Representations*, 2017. Under review.

[21] T. Tieleman and G. Hinton. Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning, 2012.

[22] Cédric Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer, Berlin, 2009.

[23] Yuhuai Wu, Yuri Burda, Ruslan Salakhutdinov, and Roger B. Grosse. On the quantitative analysis of decoder-based generative models. *CoRR*, abs/1611.04273, 2016.

[24] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. *Corr*, abs/1506.03365, 2015.

[25] Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial network. *Corr*, abs/1609.03126, 2016.

A Why Wasserstein is indeed weak

We now introduce our notation. Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a compact set (such as $[0,1]^d$ the space of images). We define $\text{Prob}(\mathcal{X})$ to be the space of probability measures over $\mathcal{X}$. We note

$$C_b(\mathcal{X}) = \{f : \mathcal{X} \to \mathbb{R},\ f \text{ is continuous and bounded}\}$$

Note that if $f \in C_b(\mathcal{X})$, we can define $\|f\|_\infty = \max_{x \in \mathcal{X}} |f(x)|$, since f is bounded. With this norm, the space $(C_b(\mathcal{X}), \|\cdot\|_\infty)$ is a normed vector space. As for any normed vector space, we can define its dual

$$C_b(\mathcal{X})^* = \{\phi : C_b(\mathcal{X}) \to \mathbb{R},\ \phi \text{ is linear and continuous}\}$$

and give it the dual norm $\|\phi\| = \sup_{f \in C_b(\mathcal{X}), \|f\|_\infty \leq 1} |\phi(f)|$.

With this definitions, $(C_b(\mathcal{X})^*, \|\cdot\|)$ is another normed space. Now let μ be a signed measure over $\mathcal{X}$, and let us define the total variation distance

$$\|\mu\|_{TV} = \sup_{A \subseteq \mathcal{X}} |\mu(A)|$$

where the supremum is taken all Borel sets in $\mathcal{X}$. Since the total variation is a norm, then if we have $\mathbb{P}_r$ and $\mathbb{P}_\theta$ two probability distributions over $\mathcal{X}$,

$$\delta(\mathbb{P}_r, \mathbb{P}_\theta) := \|\mathbb{P}_r - \mathbb{P}_\theta\|_{TV}$$

is a distance in $\text{Prob}(\mathcal{X})$ (called the total variation distance).

We can consider

$$\Phi : (\text{Prob}(\mathcal{X}), \delta) \to (C_b(\mathcal{X})^*, \|\cdot\|)$$

where $\Phi(\mathbb{P})(f) := \mathbb{E}_{x \sim \mathbb{P}}[f(x)]$ is a linear function over $C_b(\mathcal{X})$. The Riesz Representation theorem ([7], Theorem 10) tells us that Φ is an isometric immersion. This tells us that we can effectively consider $\text{Prob}(\mathcal{X})$ with the total variation distance as a subset of $C_b(\mathcal{X})^*$ with the norm distance. Thus, just to accentuate it one more time, the total variation over $\text{Prob}(\mathcal{X})$ is exactly the norm distance over $C_b(\mathcal{X})^*$.

Let us stop for a second and analyze what all this technicality meant. The main thing to carry is that we introduced a distance δ over probability distributions. When looked as a distance over a subset of $C_b(\mathcal{X})^*$, this distance gives the norm topology. The norm topology is very strong. Therefore, we can expect that not many functions $\theta \mapsto \mathbb{P}_\theta$ will be continuous when measuring distances between distributions with δ. As we will show later in Theorem 2, δ gives the same topology as the Jensen-Shannon divergence, pointing to the fact that the JS is a very strong distance, and is thus more propense to give a discontinuous loss function.

Now, all dual spaces (such as $C_b(\mathcal{X})^*$ and thus $\text{Prob}(\mathcal{X})$) have a strong topology (induced by the norm), and a weak* topology. As the name suggests, the weak* topology is much weaker than the strong topology. In the case of $\text{Prob}(\mathcal{X})$, the strong topology is given by the total variation distance, and the weak* topology is given by the Wasserstein distance (among others) [22].

B Assumption definitions

Assumption 1. *Let $g : \mathcal{Z} \times \mathbb{R}^d \to \mathcal{X}$ be locally Lipschitz between finite dimensional vector spaces. We will denote $g_\theta(z)$ it's evaluation on coordinates (z, θ). We say that g satisfies assumption 1 for a certain probability distribution p over $\mathcal{Z}$ if there are local Lipschitz constants $L(\theta, z)$ such that*

$$\mathbb{E}_{z\sim p}[L(\theta, z)] < +\infty$$

C Proofs of things

Proof of Theorem 1. Let θ and θ' be two parameter vectors in $\mathbb{R}^d$. Then, we will first attempt to bound $W(\mathbb{P}_\theta, \mathbb{P}_{\theta'})$, from where the theorem will come easily. The main element of the proof is the use of the coupling γ, the distribution of the joint $(g_\theta(Z), g_{\theta'}(Z))$, which clearly has $\gamma \in \Pi(\mathbb{P}_\theta, \mathbb{P}_{\theta'})$.

By the definition of the Wasserstein distance, we have

$$\begin{aligned} W(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) &\leq \int_{\mathcal{X}\times\mathcal{X}} \|x - y\| \, \mathrm{d}\gamma \\ &= \mathbb{E}_{(x,y)\sim\gamma}[\|x - y\|] \\ &= \mathbb{E}_z[\|g_\theta(z) - g_{\theta'}(z)\|] \end{aligned}$$

If g is continuous in θ, then $g_\theta(z) \to_{\theta\to\theta'} g_{\theta'}(z)$, so $\|g_\theta - g_{\theta'}\| \to 0$ pointwise as functions of z. Since $\mathcal{X}$ is compact, the distance of any two elements in it has to be uniformly bounded by some constant M, and therefore $\|g_\theta(z) - g_{\theta'}(z)\| \leq M$ for all θ and z uniformly. By the bounded convergence theorem, we therefore have

$$W(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) \leq \mathbb{E}_z[\|g_\theta(z) - g_{\theta'}(z)\|] \to_{\theta\to\theta'} 0$$

Finally, we have that

$$|W(\mathbb{P}_r, \mathbb{P}_\theta) - W(\mathbb{P}_r, \mathbb{P}_{\theta'})| \leq W(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) \to_{\theta\to\theta'} 0$$

proving the continuity of $W(\mathbb{P}_r, \mathbb{P}_\theta)$.

Now let g be locally Lipschitz. Then, for a given pair (θ, z) there is a constant $L(\theta, z)$ and an open set U such that $(\theta, z) \in U$, such that for every $(\theta', z') \in U$ we have

$$\|g_\theta(z) - g'_\theta(z')\| \leq L(\theta, z)(\|\theta - \theta'\| + \|z - z'\|)$$

By taking expectations and $z' = z$ we

$$\mathbb{E}_z[\|g_\theta(z) - g_{\theta'}(z)\|] \leq \|\theta - \theta'\|\mathbb{E}_z[L(\theta, z)]$$

whenever $(\theta', z) \in U$. Therefore, we can define $U_\theta = \{\theta' | (\theta', z) \in U\}$. It's easy to see that since U was open, U_θ is as well. Furthermore, by assumption 1, we can define $L(\theta) = \mathbb{E}_z[L(\theta, z)]$ and achieve

$$|W(\mathbb{P}_r, \mathbb{P}_\theta) - W(\mathbb{P}_r, \mathbb{P}_{\theta'})| \leq W(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) \leq L(\theta)\|\theta - \theta'\|$$

for all $\theta' \in U_\theta$, meaning that $W(\mathbb{P}_r, \mathbb{P}_\theta)$ is locally Lipschitz. This obviously implies that $W(\mathbb{P}_r, \mathbb{P}_\theta)$ is everywhere continuous, and by Radamacher's theorem we know it has to be differentiable almost everywhere.

The counterexample for item 3 of the Theorem is indeed Example 1. □

Proof of Corollary 1. We begin with the case of smooth nonlinearities. Since g is C^1 as a function of (θ, z) then for any fixed (θ, z) we have $L(\theta, Z) \leq \|\nabla_{\theta,x} g_\theta(z)\| + \epsilon$ is an acceptable local Lipschitz constant for all $\epsilon > 0$. Therefore, it suffices to prove

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

If H is the number of layers we know that $\nabla_z g_\theta(z) = \prod_{k=1}^H W_k D_k$ where W_k are the weight matrices and D_k is are the diagonal Jacobians of the nonlinearities. Let $f_{i:j}$ be the application of layers i to j inclusively (e.g. $g_\theta = f_{1:H}$). Then, $\nabla_{W_k} g_\theta(z) = \left(\left(\prod_{i=k+1}^H W_i D_i\right) D_k\right) f_{1:k-1}(z)$. We recall that if L is the Lipschitz constant of the nonlinearity, then $\|D_i\| \leq L$ and $\|f_{1:k-1}(z)\| \leq \|z\| L^{k-1} \prod_{i=1}^{k-1} W_i$. Putting this together,

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

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

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

finishing the proof □

Proof of Theorem 2.

1. • $(\delta(\mathbb{P}_n, \mathbb{P}) \to 0 \Rightarrow JS(\mathbb{P}_n, \mathbb{P}) \to 0)$ — Let $\mathbb{P}_m$ be the mixture distribution $\mathbb{P}_m = \frac{1}{2}\mathbb{P}_n + \frac{1}{2}\mathbb{P}$ (note that $\mathbb{P}_m$ depends on n). It is easily verified that $\delta(\mathbb{P}_m, \mathbb{P}_n) \leq \delta(\mathbb{P}_n, \mathbb{P})$, and in particular this tends to 0 (as does $\delta(\mathbb{P}_m, \mathbb{P})$). We now show this for completeness. Let μ be a signed measure, we define $\|\mu\|_{TV} = \sup_{A\subseteq\mathcal{X}} |\mu(A)|$. for all Borel sets A. In this case,

$$\begin{aligned}
\delta(\mathbb{P}_m, \mathbb{P}_n) &= \|\mathbb{P}_m - \mathbb{P}_n\|_{TV} \\
&= \|\frac{1}{2}\mathbb{P} + \frac{1}{2}\mathbb{P}_n - \mathbb{P}_n\|_{TV} \\
&= \frac{1}{2}\|\mathbb{P} - \mathbb{P}_n\|_{TV} \\
&= \frac{1}{2}\delta(\mathbb{P}_n, \mathbb{P}) \leq \delta(\mathbb{P}_n, \mathbb{P})
\end{aligned}$$

Let $f_n = \frac{d\mathbb{P}_n}{d\mathbb{P}_m}$ be the Radon-Nykodim derivative between $\mathbb{P}_n$ and the mixture. Note that by construction for every Borel set A we have $\mathbb{P}_n(A) \leq 2\mathbb{P}_m(A)$. If $A = \{f_n > 3\}$ then we get

$$\mathbb{P}_n(A) = \int_A f_n \, \mathrm{d}\mathbb{P}_m \geq 3\mathbb{P}_m(A)$$

which implies $\mathbb{P}_m(A) = 0$. This means that f_n is bounded by 3 $\mathbb{P}_m$(and therefore $\mathbb{P}_n$ and $\mathbb{P}$)-almost everywhere. We could have done this for any constant larger than 2 but for our purposes 3 will sufice.
Let $\epsilon > 0$ fixed, and $A_n = \{f_n > 1 + \epsilon\}$. Then,

$$\mathbb{P}_n(A_n) = \int_{A_n} f_n \, \mathrm{d}\mathbb{P}_m \geq (1 + \epsilon)\mathbb{P}_m(A_n)$$

Therefore,

$$\begin{aligned}
\epsilon\mathbb{P}_m(A_n) &\leq \mathbb{P}_n(A_n) - \mathbb{P}_m(A_n) \\
&\leq |\mathbb{P}_n(A_n) - \mathbb{P}_m(A_n)| \\
&\leq \delta(\mathbb{P}_n, \mathbb{P}_m) \\
&\leq \delta(\mathbb{P}_n, \mathbb{P}).
\end{aligned}$$

Which implies $\mathbb{P}_m(A_m) \leq \frac{1}{\epsilon}\delta(\mathbb{P}_n, \mathbb{P})$. Furthermore,

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

We now can see that

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

Taking limsup we get $0 \leq \limsup KL(\mathbb{P}_n \| \mathbb{P}_m) \leq \log(1 + \epsilon)$ for all $\epsilon > 0$, which means $KL(\mathbb{P}_n \| \mathbb{P}_m) \to 0$.
In the same way, we can define $g_n = \frac{d\mathbb{P}}{d\mathbb{P}_m}$, and

$$2\mathbb{P}_m(\{g_n > 3\}) \geq \mathbb{P}(\{g_n > 3\}) \geq 3\mathbb{P}_m(\{g_n > 3\})$$

meaning that $\mathbb{P}_m(\{g_n > 3\}) = 0$ and therefore g_n is bounded by 3 almost everywhere for $\mathbb{P}_n, \mathbb{P}_m$ and $\mathbb{P}$. With the same calculation, $B_n = \{g_n > 1+\epsilon\}$ and

$$\mathbb{P}(B_n) = \int_{B_n} g_n \, \mathrm{d}\mathbb{P}_m \geq (1+\epsilon)\mathbb{P}_m(B_n)$$

so $\mathbb{P}_m(B_n) \leq \frac{1}{\epsilon}\delta(\mathbb{P}, \mathbb{P}_m) \to 0$, and therefore $\mathbb{P}(B_n) \to 0$. We can now show

$$\begin{aligned} KL(\mathbb{P}\|\mathbb{P}_m) &= \int \log(g_n) \, \mathrm{d}\mathbb{P} \\ &\leq \log(1+\epsilon) + \int_{B_n} \log(g_n) \, \mathrm{d}\mathbb{P} \\ &\leq \log(1+\epsilon) + \log(3)\mathbb{P}(B_n) \end{aligned}$$

so we achieve $0 \leq \limsup KL(\mathbb{P}\|\mathbb{P}_m) \leq \log(1+\epsilon)$ and then $KL(\mathbb{P}\|\mathbb{P}_m) \to 0$. Finally, we conclude

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

- $(JS(\mathbb{P}_n, \mathbb{P}) \to 0 \Rightarrow \delta(\mathbb{P}_n, \mathbb{P}) \to 0)$ — by a simple application of the triangular and Pinsker's inequalities we get

$$\begin{aligned} \delta(\mathbb{P}_n, \mathbb{P}) &\leq \delta(\mathbb{P}_n, \mathbb{P}_m) + \delta(\mathbb{P}, \mathbb{P}_m) \\ &\leq \sqrt{\frac{1}{2}KL(\mathbb{P}_n\|\mathbb{P}_m)} + \sqrt{\frac{1}{2}KL(\mathbb{P}\|\mathbb{P}_m)} \\ &\leq 2\sqrt{JS(\mathbb{P}_n, \mathbb{P})} \to 0 \end{aligned}$$

2. This is a long known fact that W metrizes the weak* topology of $(C(\mathcal{X}), \|\cdot\|_\infty)$ on $\mathrm{Prob}(\mathcal{X})$, and by definition this is the topology of convergence in distribution. A proof of this can be found (for example) in [22].

3. This is a straightforward application of Pinsker's inequality

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

4. This is trivial by recalling the fact that δ and W give the strong and weak* topologies on the dual of $(C(\mathcal{X}), \|\cdot\|_\infty)$ when restricted to $\mathrm{Prob}(\mathcal{X})$.

□

Proof of Theorem 3. Let us define

$$\begin{aligned} V(\tilde{f}, \theta) &= \mathbb{E}_{x \sim \mathbb{P}_r}[\tilde{f}(x)] - \mathbb{E}_{x \sim \mathbb{P}_\theta}[\tilde{f}(x)] \\ &= \mathbb{E}_{x \sim \mathbb{P}_r}[\tilde{f}(x)] - \mathbb{E}_{z \sim p(z)}[\tilde{f}(g_\theta(z))] \end{aligned}$$

where $\tilde{f}$ lies in $\mathcal{F} = \{\tilde{f} : \mathcal{X} \to \mathbb{R} \text{ , } \tilde{f} \in C_b(\mathcal{X}), \|\tilde{f}\|_L \leq 1\}$ and $\theta \in \mathbb{R}^d$.

Since $\mathcal{X}$ is compact, we know by the Kantorovich-Rubenstein duality [22] that there is an $f \in \mathcal{F}$ that attains the value

$$W(\mathbb{P}_r, \mathbb{P}_\theta) = \sup_{\tilde{f} \in \mathcal{F}} V(\tilde{f}, \theta) = V(f, \theta)$$

Let us define $X^*(\theta) = \{f \in \mathcal{F} : V(f, \theta) = W(\mathbb{P}_r, \mathbb{P}_\theta)\}$. By the above point we know then that $X^*(\theta)$ is non-empty. We know that by a simple envelope theorem ([12], Theorem 1) that

$$\nabla_\theta W(\mathbb{P}_r, \mathbb{P}_\theta) = \nabla_\theta V(f, \theta)$$

for any $f \in X^*(\theta)$ when both terms are well-defined.

Let $f \in X^*(\theta)$, which we knows exists since $X^*(\theta)$ is non-empty for all θ. Then, we get

$$\begin{aligned} \nabla_\theta W(\mathbb{P}_r, \mathbb{P}_\theta) &= \nabla_\theta V(f, \theta) \\ &= \nabla_\theta [\mathbb{E}_{x \sim \mathbb{P}_r}[f(x)] - \mathbb{E}_{z \sim p(z)}[f(g_\theta(z))] \\ &= -\nabla_\theta \mathbb{E}_{z \sim p(z)}[f(g_\theta(z))] \end{aligned}$$

under the condition that the first and last terms are well-defined. The rest of the proof will be dedicated to show that

$$-\nabla_\theta \mathbb{E}_{z \sim p(z)}[f(g_\theta(z))] = -\mathbb{E}_{z \sim p(z)}[\nabla_\theta f(g_\theta(z))] \tag{5}$$

when the right hand side is defined. For the reader who is not interested in such technicalities, he or she can skip the rest of the proof.

Since $f \in \mathcal{F}$, we know that it is 1-Lipschitz. Furthermore, $g_\theta(z)$ is locally Lipschitz as a function of (θ, z). Therefore, $f(g_\theta(z))$ is locally Lipschitz on (θ, z) with constants $L(\theta, z)$ (the same ones as g). By Radamacher's Theorem, $f(g_\theta(z))$ has to be differentiable almost everywhere for (θ, z) jointly. Rewriting this, the set $A = \{(\theta, z) : f \circ g \text{ is not differentiable}\}$ has measure 0. By Fubini's Theorem, this implies that for almost every θ the section $A_\theta = \{z : (\theta, z) \in A\}$ has measure 0. Let's now fix a θ_0 such that the measure of A_{θ_0} is null (**such as when the right hand side of equation** (5) **is well defined**). For this θ_0 we have $\nabla_\theta f(g_\theta(z))|_{\theta_0}$ is well-defined for almost any z, and since $p(z)$ has a density, it is defined $p(z)$-a.e. By assumption 1 we know that

$$\mathbb{E}_{z \sim p(z)}[\|\nabla_\theta f(g_\theta(z))|_{\theta_0}\|] \leq \mathbb{E}_{z \sim p(z)}[L(\theta_0, z)] < +\infty$$

so $\mathbb{E}_{z \sim p(z)}[\nabla_\theta f(g_\theta(z))|_{\theta_0}]$ is well-defined for almost every θ_0. Now, we can see

$$\frac{\mathbb{E}_{z \sim p(z)}[f(g_\theta(z))] - \mathbb{E}_{z \sim p(z)}[f(g_{\theta_0}(z))] - \langle (\theta - \theta_0), \mathbb{E}_{z \sim p(z)}[\nabla_\theta f(g_\theta(z))|_{\theta_0}] \rangle}{\|\theta - \theta_0\|} \tag{6}$$

$$= \mathbb{E}_{z\sim p(z)}\left[\frac{f(g_\theta(z)) - f(g_{\theta_0}(z)) - \langle(\theta-\theta_0), \nabla_\theta f(g_\theta(z))|_{\theta_0}\rangle}{\|\theta-\theta_0\|}\right]$$

By differentiability, the term inside the integral converges $p(z)$-a.e. to 0 as $\theta \to \theta_0$. Furthermore,

$$\begin{aligned}\|\frac{f(g_\theta(z)) - f(g_{\theta_0}(z)) - \langle(\theta-\theta_0), \nabla_\theta f(g_\theta(z))|_{\theta_0}\rangle}{\|\theta-\theta_0\|}\| & \\ \leq \frac{\|\theta-\theta_0\|L(\theta_0,z) + \|\theta-\theta_0\|\|\nabla_\theta f(g_\theta(z))|_{\theta_0}\|}{\|\theta-\theta_0\|} & \\ \leq 2L(\theta_0,z) &\end{aligned}$$

and since $\mathbb{E}_{z\sim p(z)}[2L(\theta_0,z)] < +\infty$ by assumption 1, we get by dominated convergence that Equation 6 converges to 0 as $\theta \to \theta_0$ so

$$\nabla_\theta \mathbb{E}_{z\sim p(z)}[f(g_\theta(z))] = \mathbb{E}_{z\sim p(z)}[\nabla_\theta f(g_\theta(z))]$$

for almost every θ, and in particular when the right hand side is well defined. Note that the mere existance of the left hand side (meaning the differentiability a.e. of $\mathbb{E}_{z\sim p(z)}[f(g_\theta(z))]$) had to be proven, which we just did. □

D Energy-based GANs optimize total variation

In this appendix we show that under an optimal discriminator, energy-based GANs (EBGANs) [25] optimize the total variation distance between the real and generated distributions.

Energy-based GANs are trained in a similar fashion to GANs, only under a different loss function. They have a discriminator D who tries to minimize

$$L_D(D, g_\theta) = \mathbb{E}_{x\sim\mathbb{P}_r}[D(x)] + \mathbb{E}_{z\sim p(z)}[[m - D(g_\theta(z))]^+]$$

for some $m > 0$ and $[x]^+ = \max(0, x)$ and a generator network g_θ that's trained to minimize

$$L_G(D, g_\theta) = \mathbb{E}_{z\sim p(z)}[D(g_\theta(z))] - \mathbb{E}_{x\sim\mathbb{P}_r}[D(x)]$$

Very importantly, D is constrained to be non-negative, since otherwise the trivial solution for D would be to set everything to arbitrarily low values. The original EBGAN paper used only $\mathbb{E}_{z\sim p(z)}[D(g_\theta(z))]$ for the loss of the generator, but this is obviously equivalent to our definition since the term $\mathbb{E}_{x\sim\mathbb{P}_r}[D(x)]$ does not dependent on θ for a fixed discriminator (such as when backproping to the generator in EBGAN training) and thus minimizing one or the other is equivalent.

We say that a measurable function $D^* : \mathcal{X} \to [0, +\infty)$ is optimal for g_θ (or $\mathbb{P}_\theta$) if $L_D(D^*, g_\theta) \leq L_D(D, g_\theta)$ for all other measurable functions D. We show that such a discriminator always exists for any two distributions $\mathbb{P}_r$ and $\mathbb{P}_\theta$, and that under such a discriminator, $L_G(D^*, g_\theta)$ is proportional to $\delta(\mathbb{P}_r, \mathbb{P}_\theta)$. As a simple corollary, we get the fact that $L_G(D^*, g_\theta)$ attains its minimum value if and only if $\delta(\mathbb{P}_r, \mathbb{P}_\theta)$ is at its minimum value, which is 0, and $\mathbb{P}_r = \mathbb{P}_\theta$ (Theorems 1-2 of [25]).

Theorem 4. *Let $\mathbb{P}_r$ be a the real data distribution over a compact space $\mathcal{X}$. Let $g_\theta : \mathcal{Z} \to \mathcal{X}$ be a measurable function (such as any neural network). Then, an optimal discriminator D^* exists for $\mathbb{P}_r$ and $\mathbb{P}_\theta$, and*

$$L_G(D^*, g_\theta) = \frac{m}{2}\delta(\mathbb{P}_r, \mathbb{P}_\theta)$$

Proof. First, we prove that there exists an optimal discriminator. Let $D : \mathcal{X} \to [0, +\infty)$ be a measurable function, then $D'(x) := \min(D(x), m)$ is also a measurable function, and $L_D(D', g_\theta) \leq L_D(D, g_\theta)$. Therefore, a function $D^* : \mathcal{X} \to [0, +\infty)$ is optimal if and only if $D^{*\prime}$ is. Furthermore, it is optimal if and only if $L_D(D^*, g_\theta) \leq L_D(D, g_\theta)$ for all $D : \mathcal{X} \to [0, m]$. We are then interested to see if there's an optimal discriminator for the problem $\min_{0\leq D(x)\leq m} L_D(D, g_\theta)$.

Note now that if $0 \leq D(x) \leq m$ we have

$$\begin{aligned}
L_D(D, g_\theta) &= \mathbb{E}_{x\sim\mathbb{P}_r}[D(x)] + \mathbb{E}_{z\sim p(z)}[[m - D(g_\theta(z))]^+] \\
&= \mathbb{E}_{x\sim\mathbb{P}_r}[D(x)] + \mathbb{E}_{z\sim p(z)}[m - D(g_\theta(z))] \\
&= m + \mathbb{E}_{x\sim\mathbb{P}_r}[D(x)] - \mathbb{E}_{z\sim p(z)}[D(g_\theta(z))] \\
&= m + \mathbb{E}_{x\sim\mathbb{P}_r}[D(x)] - \mathbb{E}_{x\sim\mathbb{P}_\theta}[D(x)]
\end{aligned}$$

Therefore, we know that

$$\begin{aligned}\inf_{0\leq D(x)\leq m} L_D(D, g_\theta) &= m + \inf_{0\leq D(x)\leq m} \mathbb{E}_{x\sim\mathbb{P}_r}[D(x)] - \mathbb{E}_{x\sim\mathbb{P}_\theta}[D(x)] \\ &= m + \inf_{-\frac{m}{2}\leq D(x)\leq \frac{m}{2}} \mathbb{E}_{x\sim\mathbb{P}_r}[D(x)] - \mathbb{E}_{x\sim\mathbb{P}_\theta}[D(x)] \\ &= m + \frac{m}{2} \inf_{-1\leq f(x)\leq 1} \mathbb{E}_{x\sim\mathbb{P}_r}[f(x)] - \mathbb{E}_{x\sim\mathbb{P}_\theta}[f(x)]\end{aligned}$$

The interesting part is that

$$\inf_{-1\leq f(x)\leq 1} \mathbb{E}_{x\sim\mathbb{P}_r}[f(x)] - \mathbb{E}_{x\sim\mathbb{P}_\theta}[f(x)] = -\delta(\mathbb{P}_r, \mathbb{P}_\theta) \tag{7}$$

and there is an $f^* : \mathcal{X} \to [-1, 1]$ such that $\mathbb{E}_{x\sim\mathbb{P}_r}[f^*(x)] - \mathbb{E}_{x\sim\mathbb{P}_\theta}[f^*(x)] = -\delta(\mathbb{P}_r, \mathbb{P}_\theta)$. This is a long known fact, found for example in [22], but we prove it later for completeness. In that case, we define $D^*(x) = \frac{m}{2} f^*(x) + \frac{m}{2}$. We then have $0 \leq D(x) \leq m$ and

$$\begin{aligned}L_D(D^*, g_\theta) &= m + \mathbb{E}_{x\sim\mathbb{P}_r}[D^*(x)] - \mathbb{E}_{x\sim\mathbb{P}_\theta}[D^*(x)] \\ &= m + \frac{m}{2}\mathbb{E}_{x\sim\mathbb{P}_r}[D^*(x)] - \mathbb{E}_{x\sim\mathbb{P}_\theta}[f^*(x)] \\ &= m - \frac{m}{2}\delta(\mathbb{P}_r, \mathbb{P}_\theta) \\ &= \inf_{0\leq D(x)\leq m} L_D(D, g_\theta)\end{aligned}$$

This shows that D^* is optimal and $L_D(D^*, g_\theta) = m - \frac{m}{2}\delta(\mathbb{P}_r, \mathbb{P}_\theta)$. Furthermore,

$$\begin{aligned}L_G(D^*, g_\theta) &= \mathbb{E}_{z\sim p(z)}[D^*(g_\theta(z))] - \mathbb{E}_{x\sim\mathbb{P}_r}[D^*(x)] \\ &= -L_D(D^*, g_\theta) + m \\ &= \frac{m}{2}\delta(\mathbb{P}_r, \mathbb{P}_g)\end{aligned}$$

concluding the proof.

For completeness, we now show a proof for equation (7) and the existence of said f^* that attains the value of the infimum. Take $\mu = \mathbb{P}_r - \mathbb{P}_\theta$, which is a signed measure, and (P, Q) its Hahn decomposition. Then, we can define $f^* := \mathbb{1}_Q - \mathbb{1}_P$. By construction, then

$$\begin{aligned}EE_{x\sim\mathbb{P}_r}[f^*(x)] - \mathbb{E}_{x\sim\mathbb{P}_\theta}[f^*(x)] &= \int f^* \,\mathrm{d}\mu = \mu(Q) - \mu(P) \\ &= -(\mu(P) - \mu(Q)) = -\|\mu\|_{TV} \\ &= -\|\mathbb{P}_r - \mathbb{P}_\theta\|_{TV} \\ &= -\delta(\mathbb{P}_r, \mathbb{P}_\theta)\end{aligned}$$

Furthermore, if f is bounded between -1 and 1, we get

$$\begin{aligned}
|\mathbb{E}_{x\sim\mathbb{P}_r}[f(x)] - \mathbb{E}_{x\sim\mathbb{P}_\theta}[f(x)]| &= |\int f \,\mathrm{d}\mathbb{P}_r - \int f \,\mathrm{d}\mathbb{P}_\theta| \\
&= |\int f \,\mathrm{d}\mu| \\
&\leq \int |f| \,\mathrm{d}|\mu| \leq \int 1 \,\mathrm{d}|\mu| \\
&= |\mu|(\mathcal{X}) = \|\mu\|_{TV} = \delta(\mathbb{P}_r, \mathbb{P}_\theta)
\end{aligned}$$

Since δ is positive, we can conclude $\mathbb{E}_{x\sim\mathbb{P}_r}[f(x)] - \mathbb{E}_{x\sim\mathbb{P}_\theta}[f(x)] \geq -\delta(\mathbb{P}_r, \mathbb{P}_\theta)$. $\square$

E Generator's cost during normal GAN training

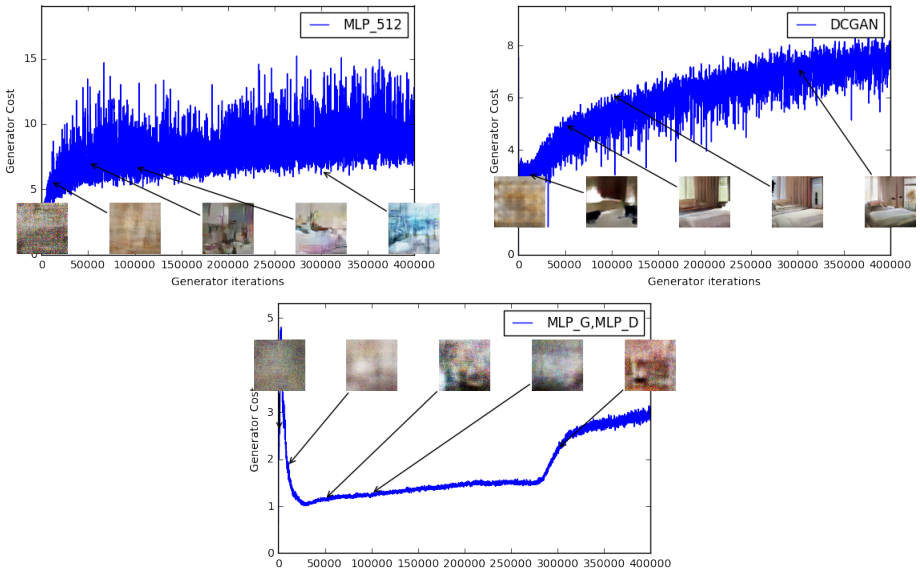

Figure 8: 在正常*GAN*训练过程中，用于*MLP*发生器*(*左上*)*和*DCGAN*发生器*(*右上*)*。两者都有一个*DCGAN*鉴别器。**两条曲线误差越来越大**。*DCGAN*的样本变得更好了但发电机的成本增加，指向样品质量与损失之间无显著相关性。底部*: MLP*同时具有生成器和鉴别器。曲线是这样的上下不分样品质量。所有培训曲线通过与*Figure 3*相同的中值过滤器。

F Sheets of samples

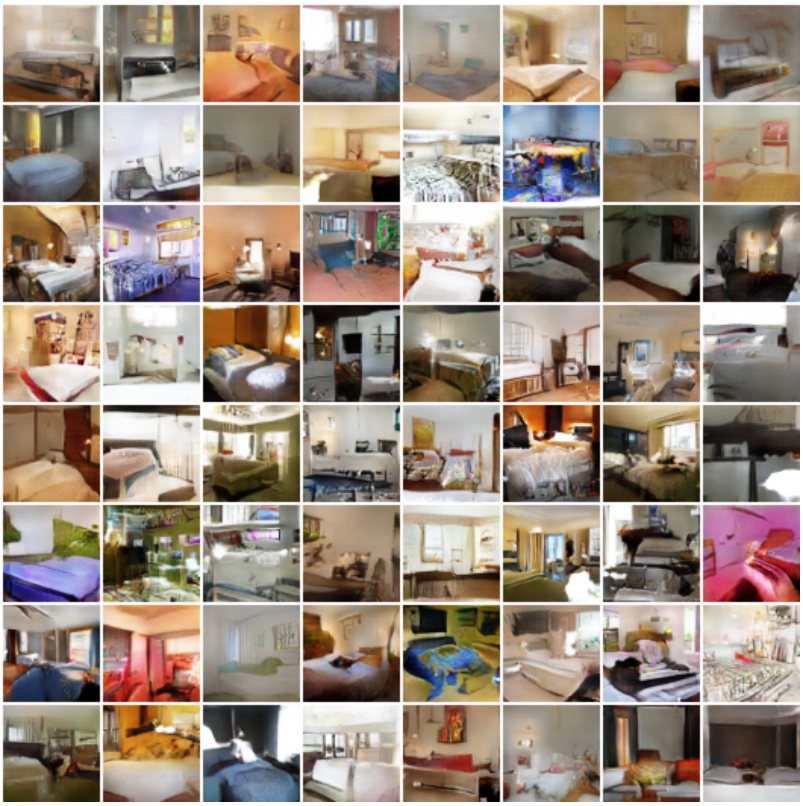

Figure 9: *WGAN*算法:发生器与批判器都是*dcgan*。

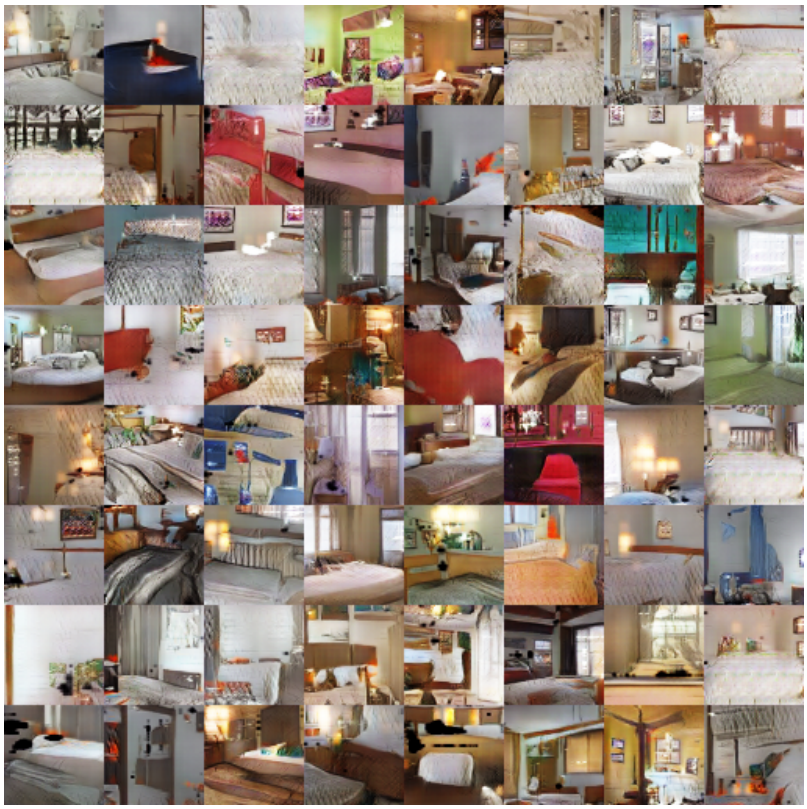

Figure 10: 标准*GAN*程序:生成器和鉴别器都是*dcgan*。

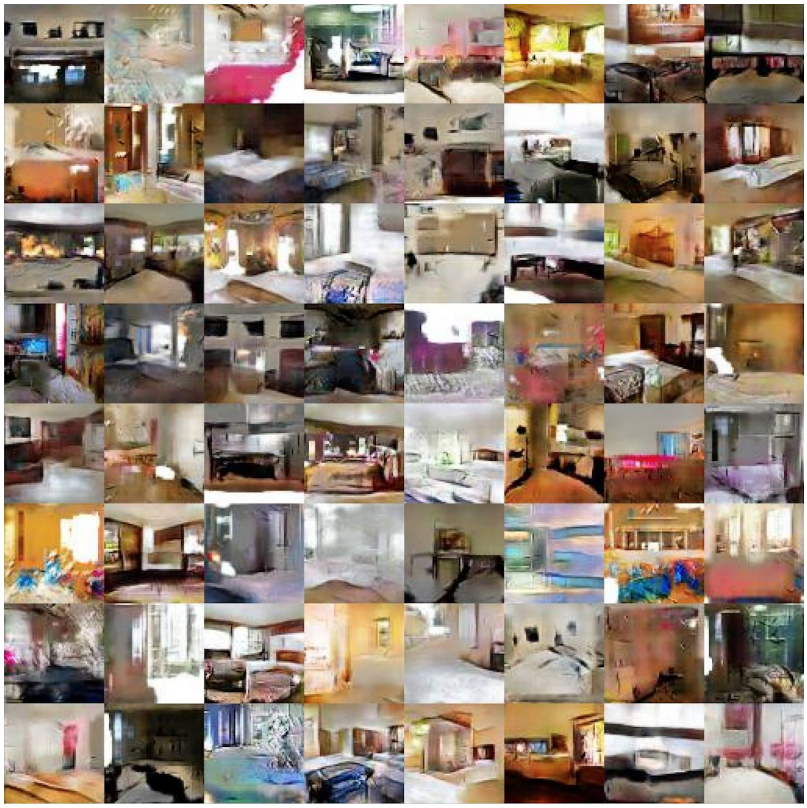

*Figure 11: WGAN*算法:生成器是一个没有*DCGAN*的批规范和恒定的过滤器大小。评论家是一个*DCGAN*。

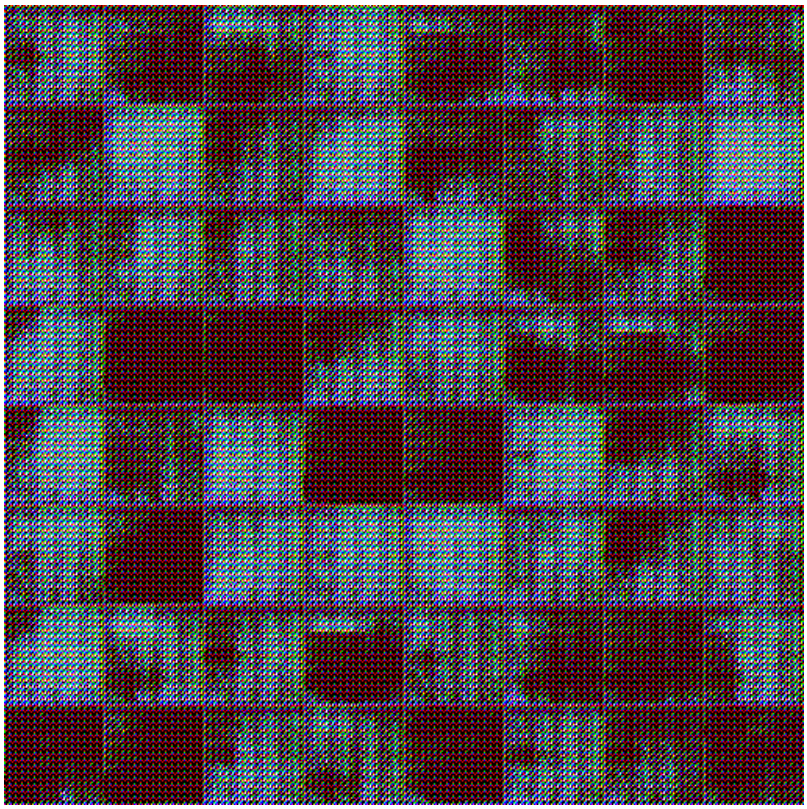

Figure 12: 标准*GAN*程序:发生器是无*DCGAN*的批规范和恒定的过滤器大小。鉴别器是一个*DCGAN*。

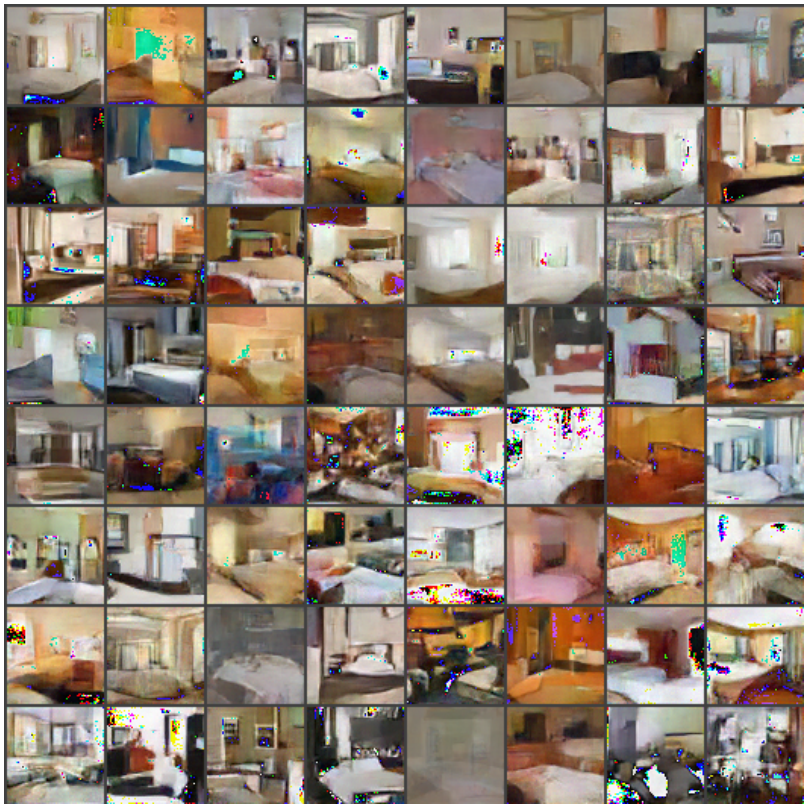

*Figure 13: WGAN*算法:生成器是一个带有*4*个隐藏的*MLP 512*个单位的层，评论家是一个*DCGAN*。

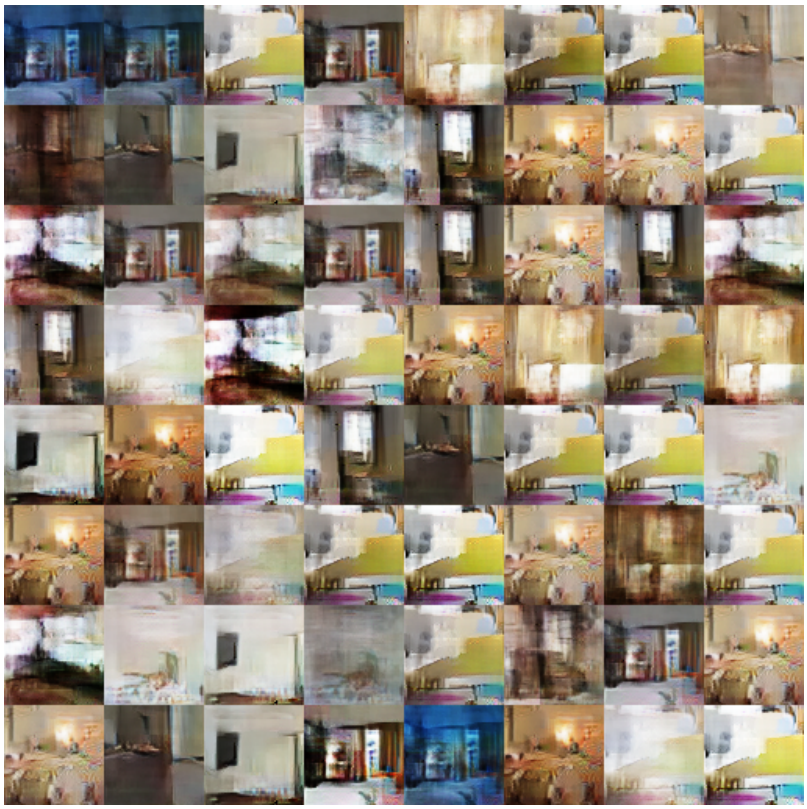

Figure 14: 标准*GAN*程序:发生器是一个*MLP*与*4*个隐藏层*512*个单元，鉴别器为*DCGAN*。