GAN训练：理论篇

GAN训练篇

1.GAN“不好训” 的理论分析

1. **[GAN, NIPS]** Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y.. Generative adversarial nets. In *Advances in neural information processing systems* (NIPS 2014).
2. **[pre_WGAN, ICLR]** Martín Arjovsky, Léon Bottou. Towards Principled Methods for Training Generative Adversarial Networks. ICLR (2017)

2.从理论分析出发的解决方法

3. **[WGAN, ICML]** Arjovsky M, Chintala S, Bottou L. Wasserstein GAN. *International Conference on Machine Learning* (ICML 2017).
4. **[f-GAN, NIPS]** Nowozin, S., Cseke, B., & Tomioka, R. f-gan: Training generative neural samplers using variational divergence minimization. *In Advances in Neural Information Processing Systems* (NIPS 2016).
5. **[EBGAN, ICLR]** Zhao J, Mathieu M, LeCun Y. Energy-based generative adversarial network. *International Conference on Learning Representations* (ICLR 2016).
6. **[BEGAN, ARXIV]** Berthelot D, Schumm T, Metz L. BeGAN: Boundary equilibrium generative adversarial networks. *arXiv*, 2017.
7. **[DCGAN, ICLR]** Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. *International Conference on Learning Representations* (ICLR 2015).

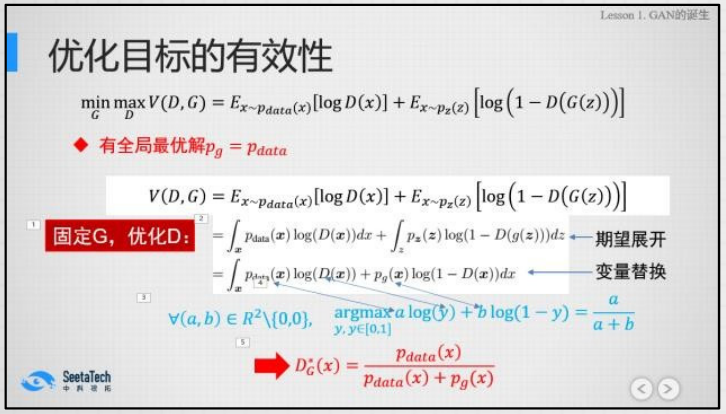
3.行之有效的启发式的解决方法

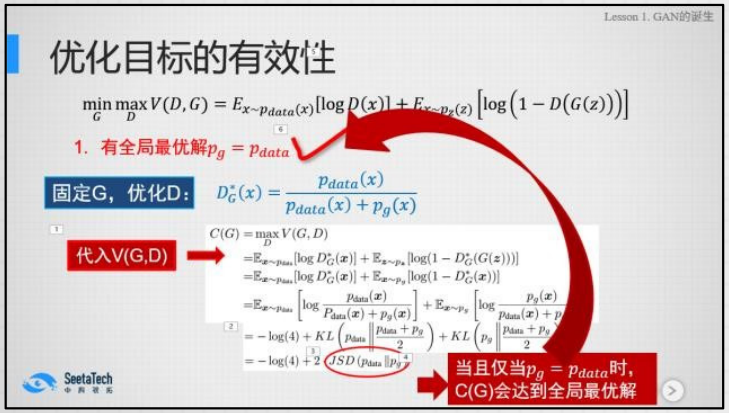
8. **[for_GAN, NIPS]** Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., techniques for training gans. *In Advances in Neural Information Processing Systems (NIPS)*, 2016.
9. **[SN-GANs, ICLR]** Miyato T, Kataoka T, Koyama M, et al. Spectral normalization for generative adversarial networks. *International Conference on Learning Representations* (ICLR 2018).

回顾GAN

★ 原始优化问题：

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}\left[\log\left(1-D\left(G(z)\right)\right)\right]$$





★ Step-1. G固定，以最大化V(D,G)为训练目标，训练D；

$$\max_D V(D,G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}\left[\log\left(1-D\left(G(z)\right)\right)\right]$$

$$D_G^*(x) = \frac{p_{data}(x)}{p_{data}(x)+p_g(x)}, \qquad V(G,D^*) = -\log 4 + 2JSD(p_{data}\|p_g)$$

等价于：

$$\min_D -V(D,G) = E_{x\sim p_{data}(x)}[-\log D(x)] + E_{z\sim p_z(z)}\left[-\log\left(1-D\left(G(z)\right)\right)\right]$$

$$D_G^*(x) = \frac{p_{data}(x)}{p_{data}(x)+p_g(x)}, \qquad -V(G,D^*) = \log 4 - 2JSD(p_{data}\|p_g)$$

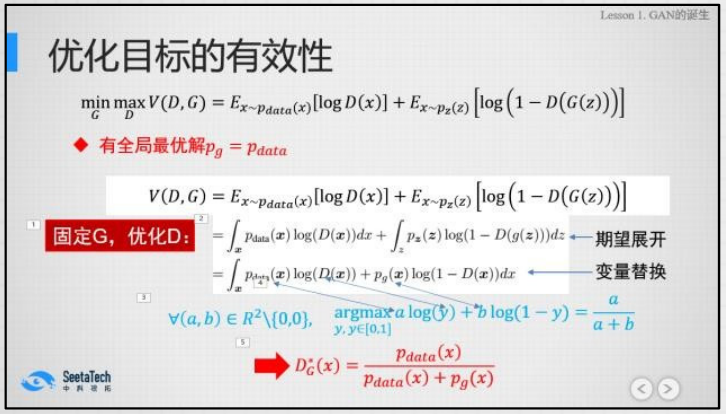
★ Step-2. D固定为D*，以最小化V(D,G)为训练目标，训练G；

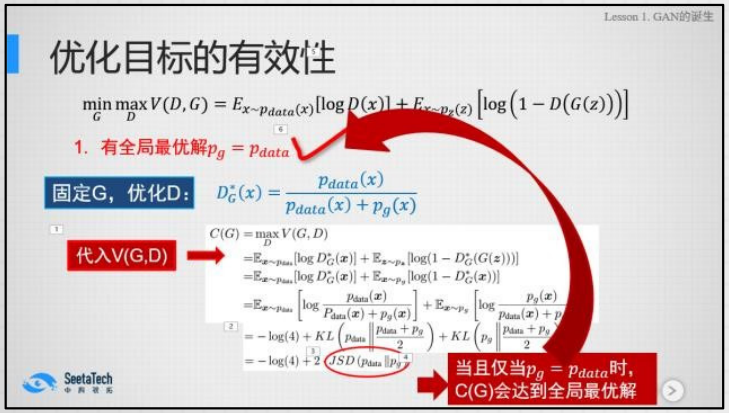
$$\min_G V(D^*,G) = E_{x\sim p_{data}(x)}[\log D^*(x)] + E_{z\sim p_z(z)}\left[\log\left(1-D^*\left(G(z)\right)\right)\right]$$

回顾GAN

★ 原始优化问题：

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}\left[\log\left(1 - D(G(z))\right)\right]$$





★ Step-1. G固定，以最大化V(D,G)为训练目标，训练D；

$$\max_D V(D,G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}\left[\log\left(1 - D(G(z))\right)\right]$$

$$D_G^*(x) = \frac{p_{data}(x)}{p_{data}(x) + p_g(x)}, \qquad V(G,D^*) = -\log 4 + 2JSD(p_{data}\|p_g)$$

等价于：

$$\min_D -V(D,G) = E_{x\sim p_{data}(x)}[-\log D(x)] + E_{z\sim p_z(z)}\left[\log\left(1 - D(G(z))\right)\right]$$

$$D_G^*(x) = \frac{p_{data}(x)}{p_{data}(x) + p_g(x)}, \qquad -V(G,D^*) = \log 4 - 2JSD(p_{data}\|p_g)$$

矛盾点1：D的理论最优 vs 实际最优。

★ Step-2. D固定为D*，以最

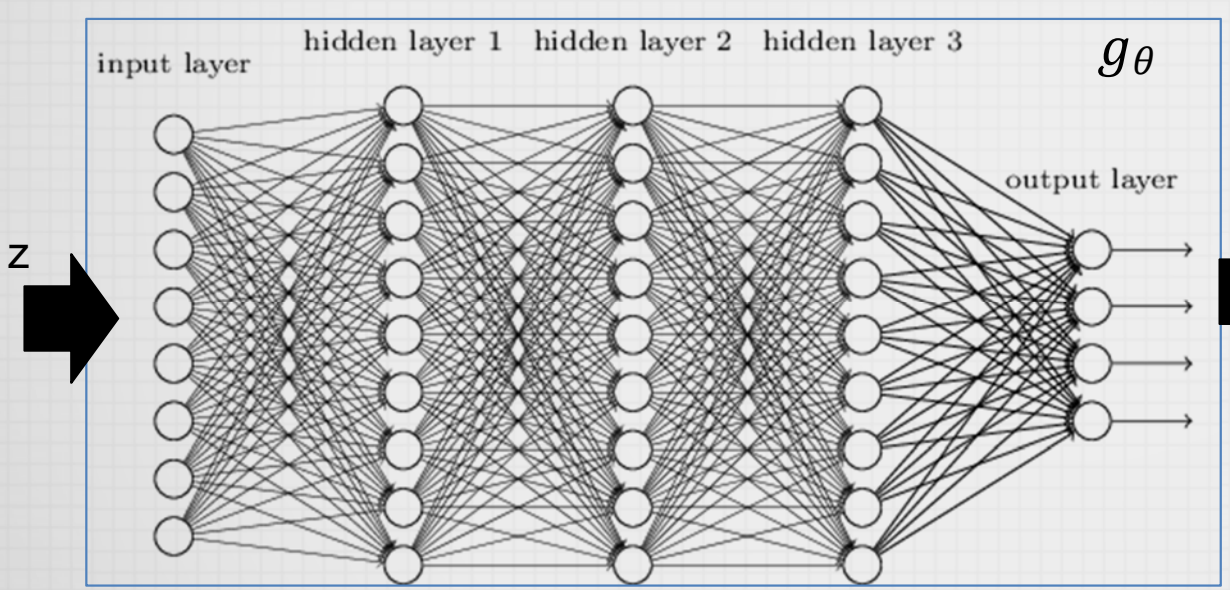
矛盾点2：D越好，G可能变好，也可能不变好-_-

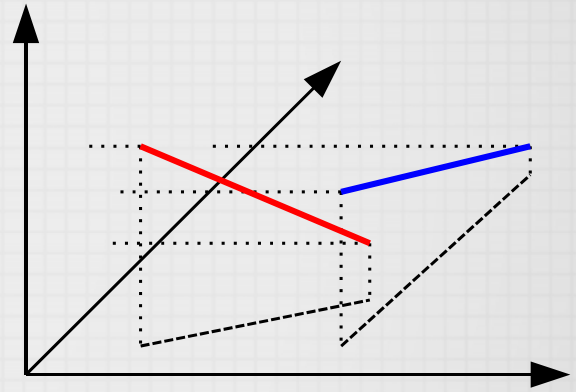
$$\min_G V(D^*,G) = E_{x\sim p_{data}(x)}[\log D^*(x)] + E_{z\sim p_z(z)}\left[\log\left(1 - D^*(G(z))\right)\right]$$

问题分析

问题1. 理论最优 vs 实际最优：D训练的时候实际最优总可以达到error=0，而理论最优是-V(G,D*)...丶(一_一)ノ

一个结论：若g是由仿射变换、pointwise的非线性操作等构成的函数，实现从某个随机变量z到X空间的映射，那么可以证明g(z)会被限定在最多z的维度的流形上，因而，如果z的维度小于X的维度，那么g(z)在X上的测度为0；





Pr Pg's support M & P, disjoint

➡ 存在准确率为1的判别器D*: MuP → [0,1]

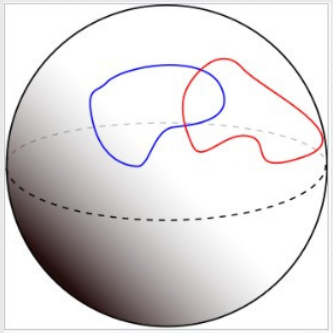
又一个结论：如果两个分布Pr, Pg的support是在两个disjoint的流形上，分别记为M和P，那么一定存在一个准确率为1的分类器，可以准确区分M和P中的任意样本。(Urysohn theorem)

问题分析

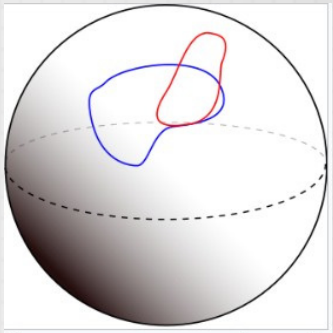
1.disjoint；2.general

问题1. 理论最优 vs 实际最优：D训练的时候实际最优总可以达到error=0，而理论最优是-V(G,D*)...ヽ(ー_ー)ノ

概念1：transversallity (横截)：交集中点在M上的切空间+在P上的切空间，可以得到整个$F=R^d$上的切空间。换句话说，就是M上的切空间和P上的切空间不重合。这个概念是与正切相对应。如下图：



Transverse curves on the surface of a sphere



Non-transverse curves on the surface of a sphere

概念2：perfectly align：如果存在有一个M和P的交集，在这个交集里存在一个点，不满足transversally的条件，就认为M和P是perfectly align的。

Theorem2.2: 如果Pr, Pg的support是在两个没有perfectly align的流形M和P上，即便它们不是disjoint的，但M和P都不是full dimension的，同时Pr，Pg在各自的流形上是连续的，那么就可以得到此时存在一个最优的判别器D*，准确率为1。(同时，此时，因为其准确率为1，导致其数值为常数，因而梯度为0.)

问题2

问题2. 训练完D达到“最优”，再训练G时，G有可能变好，也有可能不变好 ....

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}\left[\log\left(1 - D(G(z))\right)\right]$$

结论1： 若$\|D - D^*\| < \epsilon$, $E_{z\sim p(z)}[\|\nabla_\theta g_\theta(z)\|_2^2] \le M^2$,并定义$\|D\| = \sup_x |D(x)| + \|\nabla_x D(x)\|_2$

$$\Rightarrow \left\|\nabla_\theta E_{z\sim p(z)}\left[\log\left(1 - D\left(g_\theta(z)\right)\right)\right]\right\|_2 < M\frac{\epsilon}{1-\epsilon} \Rightarrow \lim_{\|D-D^*\|\to 0} \nabla_\theta E_{z\sim p(z)}\left[\log\left(1 - D\left(g_\theta(z)\right)\right)\right] = 0$$

$$\begin{aligned}
\|\nabla_\theta \mathbb{E}_{z\sim p(z)}[\log(1 - D(g_\theta(z)))]\|_2^2 &\le \mathbb{E}_{z\sim p(z)}\left[\frac{\|\nabla_\theta D(g_\theta(z))\|_2^2}{|1 - D(g_\theta(z))|^2}\right] \\
&\le \mathbb{E}_{z\sim p(z)}\left[\frac{\|\nabla_x D(g_\theta(z))\|_2^2 \|J_\theta g_\theta(z)\|_2^2}{|1 - D(g_\theta(z))|^2}\right] \\
&< \mathbb{E}_{z\sim p(z)}\left[\frac{\left(\|\nabla_x D^*(g_\theta(z))\|_2 + \epsilon\right)^2 \|J_\theta g_\theta(z)\|_2^2}{\left(|1 - D^*(g_\theta(z))| - \epsilon\right)^2}\right] \\
&= \mathbb{E}_{z\sim p(z)}\left[\frac{\epsilon^2 \|J_\theta g_\theta(z)\|_2^2}{(1-\epsilon)^2}\right] \\
&\le M^2 \frac{\epsilon^2}{(1-\epsilon)^2}
\end{aligned}$$

$Var(X) = E(X^2) - [E(x)]^2 \ge 0$

$\|\cdot\|$定义

$\|D - D^*\| < \epsilon$

D^*为常数，且$D^*\left(g_\theta(z)\right) = 0$

问题2

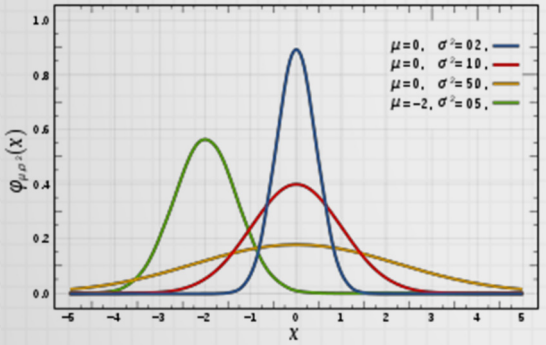
1.梯度消失；2.梯度不稳定

问题2. 训练完D达到“最优”，再训练G时，G有可能变好，也有可能不变好 ....

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}\left[\log\left(1 - D(G(z))\right)\right]$$

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}\left[-\log\left(D(G(z))\right)\right]$$

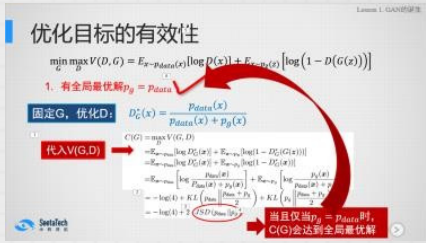
$$E_{z\sim p(z)}\left[-\log D^*(g_\theta(z))\right] = E_{z\sim p(z)}\left[\log\left(\frac{1 - D^*(g_\theta(z))}{D^*(g_\theta(z))} * \frac{1}{\left[1 - D^*(g_\theta(z))\right]}\right)\right]$$



$$= \mathrm{E}_z\left[\log\frac{p_g(g_\theta(z))}{p_r(g_\theta(z))}\right] - \mathrm{E}_z\left[\log\left(1 - D^*(g(z))\right)\right]$$

$$= \mathrm{E}_{x\sim p}\left[\log\frac{p_g}{p_r}\right] - \mathrm{E}_{x\sim p}\left[\log(1 - D^*)\right]$$

$$= KL(p_g, p_r) - 2JSD(p_g, p_r) + \log 4$$



$$-\log 4 + 2JSD(p_g, p_r)$$

结论2：在与前面定理类似的容易满足的条件下，优化目标对于生成器参数的梯度的期望有无限大的期望与方差。

总结

问题1. D理论最优vs实际最优，无法提供梯度帮助：

在多数情况(因为条件很容易满足，低维流形，disjoint)下，D总能完美分类----> D的梯度为0 --------> D无法帮助G去训练

问题2：训练不稳定

1. 训练G时的目标函数的梯度会消失
2. 训练G时的目标函数的梯度的方差无限大

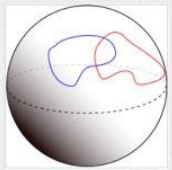
GAN问题回顾

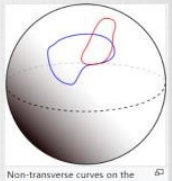
问题分析

1.disjoint；2.general

问题1. 理论最优 vs 实际最优：D训练的时候实际最优总可以达到error=0，而理论最优是-V(G,D*)...ヽ(ー_ー)ノ

概念1：transversallity (横截)：交集中点在M上的切空间+在P上的切空间，可以得到整个F=Rd上的切空间。换句话说，就是M上的切空间和P上的切空间不重合。这个概念是与正切相对应。如下图：





概念2：perfectly align：如果存在有一个M和P的交集，在这个交集里存在一个点，不满足transversally的条件，就认为M和P是perfectly align的。

Theorem2.2: 如果Pr, Pg的support是在两个没有perfectly align的流形M和P上，即便它们不是disjoint的，但M和P都不是full dimension的，同时Pr，Pg在各自的流形上是连续的，那么就可以得到此时存在一个最优的判别器D*，准确率为1。(同时，此时，因为其准确率为1，导致其数值为常数，因而梯度为0.)

SeetaTech 中科视拓

问题2

问题2. 训练完D达到“最优”，再训练G时，G有可能变好，也有可能不变好 ....

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}\left[\log\left(1-D(G(z))\right)\right]$$

结论1： 若$\|D-D^*\|<\epsilon$, $E_{z\sim p(z)}[\|\nabla_\theta g_\theta(z)\|_2^2] \le M^2$,并定义$\|D\| = \sup_x |D(x)| + \|\nabla_x D(x)\|_2$

$$\Rightarrow \left\|\nabla_\theta E_{z\sim p(z)}\left[\log\left(1-D(g_\theta(z))\right)\right]\right\|_2 < M\frac{\epsilon}{1-\epsilon} \quad \Rightarrow \lim_{\|D-D^*\|\to 0} \nabla_\theta E_{z\sim p(z)}\left[\log\left(1-D(g_\theta(z))\right)\right] = 0$$

$$\begin{aligned}
\|\nabla_\theta \mathrm{E}_{z\sim p(z)}[\log(1-D(g_\theta(z)))]\|_2^2 &\le \mathrm{E}_{z\sim p(z)}\left[\frac{\|\nabla_\theta D(g_\theta(z))\|_2^2}{|1-D(g_\theta(z))|^2}\right] \\
&\le \mathrm{E}_{z\sim p(z)}\left[\frac{\|\nabla_x D(g_\theta(z))\|_2^2 \|J_\theta g_\theta(z)\|_2^2}{|1-D(g_\theta(z))|^2}\right] \\
&< \mathrm{E}_{z\sim p(z)}\left[\frac{(\|\nabla_x D^*(g_\theta(z))\|_2+\epsilon)^2 \|J_\theta g_\theta(z)\|_2^2}{(|1-D^*(g_\theta(z))|-\epsilon)^2}\right] \\
&= \mathrm{E}_{z\sim p(z)}\left[\frac{\epsilon^2\|J_\theta g_\theta(z)\|_2^2}{(1-\epsilon)^2}\right] \\
&\le M^2\frac{\epsilon^2}{(1-\epsilon)^2}
\end{aligned}$$

$Var(X) = E(X^2) - [E(x)]^2 \ge 0$

$\|\cdot\|$定义

$\|D-D^*\|<\epsilon$

D^*为常数，且$D^*(g_\theta(z)) = 0$

SeetaTech 中科视拓

问题2

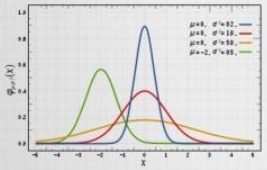
1.梯度消失；2.梯度不稳定

问题2. 训练完D达到“最优”，再训练G时，G有可能变好，也有可能不变好 ....

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}\left[\log\left(1-D(G(z))\right)\right]$$

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}\left[-\log\left(D(G(z))\right)\right]$$

$$E_{z\sim p(z)}[-\log D^*(g_\theta(z))] = E_{z\sim p(z)}\left[\log\left(\frac{1-D^*(g_\theta(z))}{D^*(g_\theta(z))} * \frac{1}{[1-D^*(g_\theta(z))]}\right)\right]$$



$$\begin{aligned}
&= \mathrm{E}_z\left[\log\frac{p_g(g_\theta(z))}{p_r(g_\theta(z))}\right] - \mathrm{E}_z\left[\log\left(1-\mathrm{D}^*(\mathrm{g}(z))\right)\right] \\
&= \mathrm{E}_{\mathrm{x}\sim p_g}\left[\log\frac{p_g}{p_r}\right] - \mathrm{E}_{\mathrm{x}\sim p_g}[\log(1-D^*)] \\
&= KL(p_g,p_r) - 2JSD(p_g,p_r) + \log 4
\end{aligned}$$

$-\log 4 + 2JSD(p_g,p_r)$

结论2：在与前面定理类似的容易满足的条件下，优化目标对于生成器参数的梯度的期望有无限大的期望与方差。

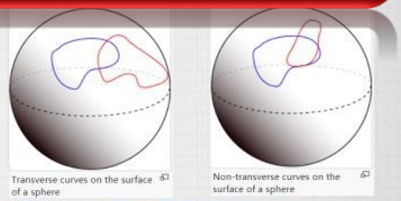
SeetaTech 中科视拓

解决方法

问题分析

问题1. 理论最优 vs 实际最优： D训练的时候实际最优总可以达到error=0，而理论最优是-V(G,D*)...丶(一_一)

概念1：transversallity (横截)：交集中点在M上的切空间+在P上的切空间，可以得到整个F=Rd上的切空间。换句话说，就是M上的切空间和P上的切空间不重合。这个概念是与下切相对应，如下图



概念2：perfectly align：如果存在有一个M和P的交集，在这个交集里存在一个点，不满足transversally的条件，就认为M和P是perfectly align的。

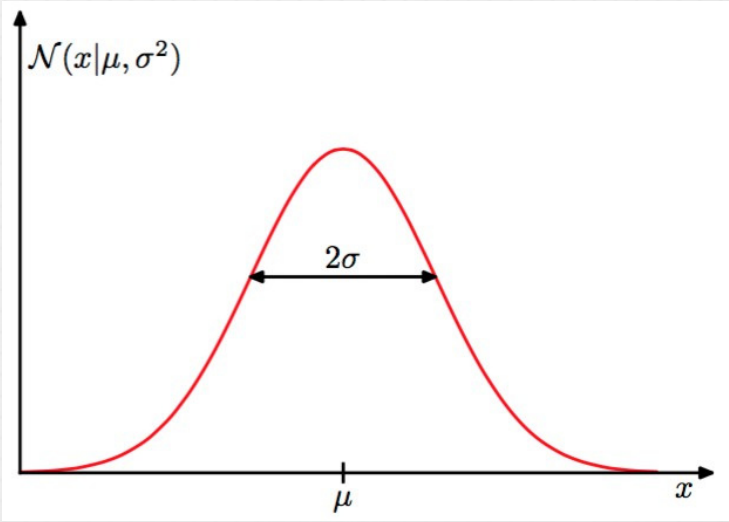
Theorem2.2: 如果Pr, Pg的support是在两个没有perfectly align的流形M和P上，即便它们不是disjoint的，但M和P都不是full dimension的，同时Pr，Pg在各自的流形上是连续的，那么就可以得到此时存在一个最优的判别器D*，准确率为1。(同时，此时，因为其准确率为1，是常数使得梯度为0，因而梯度为0.)

SeetaTech 中科视拓

例：
[23]: 数据本身:[0,1];
噪声方差0.1

解决方法1：扩大支撑集范围

→ 添加噪声项

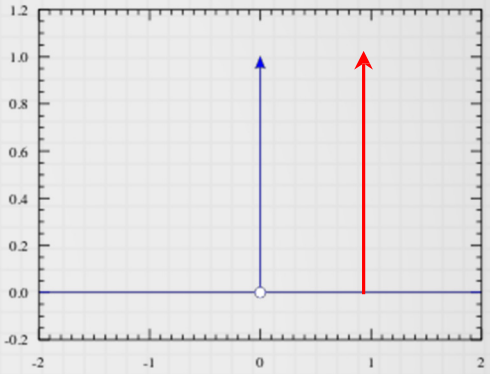


优点：不需要担心D的优化程度

缺点：噪声？？Are you kidding me??

不同距离度量的分析

- $p_0(0,z);\ g_\theta(\theta,z)$



- KL距离：不对称

$$KL(p_0|p_\theta)=\sum_x p_0\log\frac{p_0}{p_\theta}=0_{x=\theta\neq 0}+\infty_{x=0\neq\theta}=\infty\ (\theta\neq 0)$$

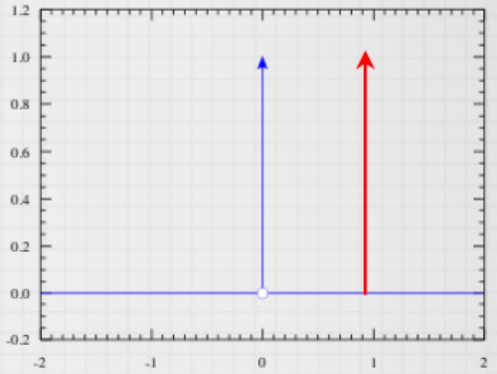
$$KL(p_\theta|p_0)=\sum_x p_\theta\log\frac{p_\theta}{p_0}=\infty_{x=\theta\neq 0}+0_{x=0\neq\theta}=\infty\ (\theta\neq 0)$$

$$\Rightarrow KL(p_\theta|p_0)=KL(p_0|p_\theta)=\begin{cases}\infty,\theta\neq 0\\ 0,\theta=0\end{cases}$$

⇒非有效距离。也无法优化。不可用！

不同距离度量的分析

- $p_0(0,z)$; $g_\theta(\theta,z)$



- JS距离：具有对称性

$$JS(p_0|p_\theta) = \frac{1}{2}KL\left(p_0\middle|\frac{p_0+p_\theta}{2}\right)+\frac{1}{2}KL\left(p_\theta\middle|\frac{p_0+p_\theta}{2}\right)$$

$$= \frac{1}{2}\sum_x p_0 \log\frac{p_0 * 2}{p_0+p_\theta} + \frac{1}{2}\sum_x p_\theta \log\frac{p_\theta * 2}{p_0+p_\theta}$$

$$= \log 2 + \frac{1}{2}\sum_x p_0 \log p_0 + \frac{1}{2}\sum_x p_\theta \log p_\theta - \frac{1}{2}\sum_x (p_0+p_\theta)\log(p_0+p_\theta)$$

$$= \log 2 + \frac{1}{2}\sum_x p_0 \log p_0 + \frac{1}{2}\sum_x p_\theta \log p_\theta - \frac{1}{2}\sum_{x\in S(p_0)} p_0 \log p_0 - \frac{1}{2}\sum_{x\in S(p_\theta)} p_\theta \log p_\theta$$

$$= \log 2\ (\theta \neq 0)$$

$$\Rightarrow \begin{cases} \log 2, \theta \neq 0 \\ 0, \theta = 0 \end{cases}$$

⇒是有效距离。但无法优化。不可用！

不同距离度量的分析

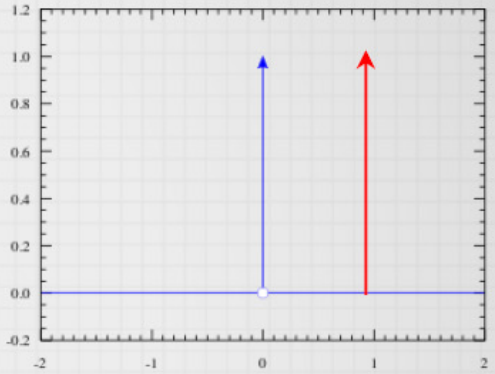
- $p_0(0,z)$; $g_\theta(\theta,z)$

- TV距离(Total Variance)： 两个分布对于同一个事件给出的概率值的最大差值。

$$TV(p_0|p_\theta) = \sup|p_0(A) - p_\theta(A)|$$

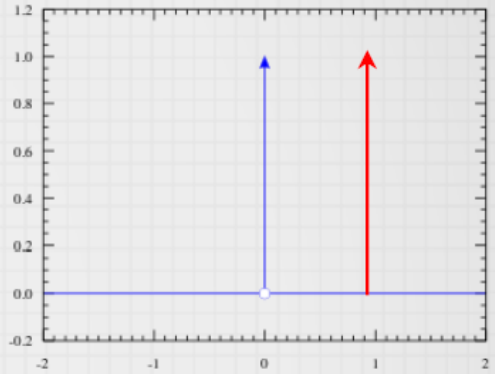
$$= \begin{cases} 1, \theta \neq 0 \\ 0, \theta = 0 \end{cases}$$

⇒是有效距离。但无法优化。不可用！



不同距离度量的分析

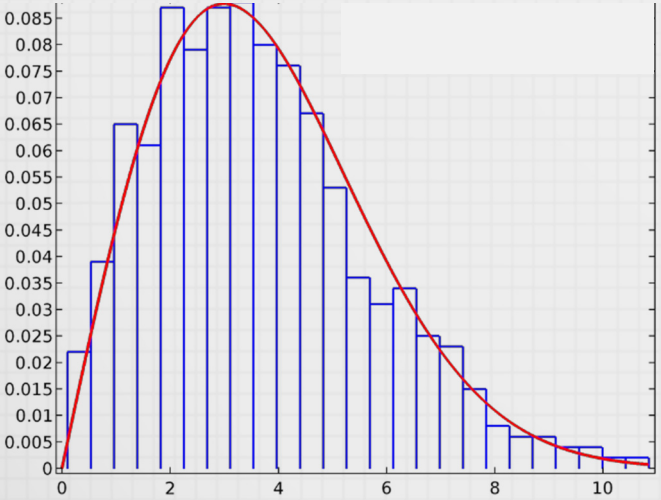
- $p_0(0,z)$; $g_\theta(\theta,z)$



- Wasserstein距离(Earth-Mover distance)：

① which-to-which; ② how much;

$$W(p_0,p_\theta)=\inf_{\gamma\sim\Pi(p_0,p_\theta)}E_{(x,y)\sim\gamma}[\|x-y\|]$$



$$p(x,y)=p(x)p(y)$$

$$\Rightarrow W(p_0,p_\theta)=\inf\sum_{x,y}p(x,y)\|x-y\|$$

$$=\inf\sum_x\sum_y p(x)p(y)\|x-y\|$$

$$=\inf\|0-\theta\|=\|\theta\|=|\theta|$$

$\Rightarrow$ 有效距离。可以优化。可用！

Wasserstein距离

- **Wasserstein距离的连续性：**
  - P_r表示空间X上的某个固定分布，z是空间Z上的随机变量，比如是高斯分布的随机变量；那么假设$g_\theta(z)$是将随机变量z映射到X空间上的函数，其中θ是参数。用P_θ表示$g_\theta(z)$的分布。那么若$g_\theta(z)$是关于θ的连续函数，则wassertein距离$W(P_r, P_\theta)$也是关于θ的连续函数。

证明：$\because W(p_0, p_\theta) = \inf_{\gamma \sim \Pi(p_0, p_\theta)} E_{(x,y)\sim\gamma}[\|x-y\|]$

$$\Longrightarrow W(P_\theta, P_{\theta'}) \le \sum_{x,y} p(x,y)\|x-y\| = E_{(x,y)\sim\gamma}[\|x-y\|] = E_z[\|g_\theta(z) - g_{\theta'}(z)\|]$$

$\because \lim_{\theta\to\theta'} g_\theta(z) = g_{\theta'}(z)$

$$\Longrightarrow \lim_{\theta\to\theta'} W(P_\theta, P_{\theta'}) \le \lim_{\theta\to\theta'} E_z[\|g_\theta(z) - g_{\theta'}(z)\|] = 0$$

$\because$ distance measure's Triangle Inequality

$$\Longrightarrow |W(P_r, P_\theta) - W(P_r, P_{\theta'})| \le W(P_\theta, P_{\theta,})$$

$$\Longrightarrow \lim_{\theta\to\theta'} W(P_r, P_\theta) = W(P_r, P_{\theta'})$$ 。得证。

Wasserstein距离

- **Wasserstein距离的近似形式：**
  - Kantorovich-Rubinstein duality：

$$W(P_r, P_\theta) = \sup_{\|f\|_L \le 1} E_{x\sim P_r}[f(x)] - E_{x\sim P_\theta}[f(x)] = \frac{1}{K} \sup_{\|f\|_L \le K} E_{x\sim P_r}[f(x)] - E_{x\sim P_\theta}[f(x)]$$

注：以上是对所有能满足$\|f\|_L \le 1$的函数而言。$\Rightarrow$ $W(P_r, P_\theta) = \max_{\|f\|_L \le 1} E_{x\sim P_r}[f(x)] - E_{x\sim P_\theta}[f(x)]$

➡ 问题：如何找到max问题的解函数f?

用w表示f函数的参数，那么当f的参数w所在的空间是compact的，即闭合且有界的时候，所有以这些w为参数的参数都是满足$\|f\|_L \le K$的条件。

➡ 问题：要如何保证参数空间是compact的，即闭合且有界的？

参数截断。

➡ [前述条件] Pr表示空间X上的某个固定分布，z是空间Z上的随机变量，比如是高斯分布的随机变量；那么假设$g_\theta(z)$是将随机变量z映射到X空间上的函数，其中θ是参数。用P_θ表示$g_\theta(z)$的分布。那么

当找到上述max问题的解函数f后，有：$\nabla_\theta W(P_r, P_\theta) = -E_{z\sim p(z)}[\nabla_\theta f(g_\theta(z))]$

WGAN

- **WGAN优化问题：**

$$\min_{\theta}\max_{w} E_{x\sim P_r}[f_w(x)] - E_{z\sim p_z}[f_w(g_\theta(z))]$$

类比原始GAN：

$$\min_{\theta}\max_{w} E_{x\sim P_r}[\log D_w(x)] - E_{z\sim p_z}[\log D_w(g_\theta(z))]$$

变化：

1 没有log；

2 要约束参数w在一个compact的空间中。（做法：参数截断）

3 不要求f_w的取值范围。

Algorithm 1 WGAN, our proposed algorithm. All experiments in the paper used the default values $\alpha = 0.00005$, $c = 0.01$, $m = 64$, $n_{\text{critic}} = 5$.

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.

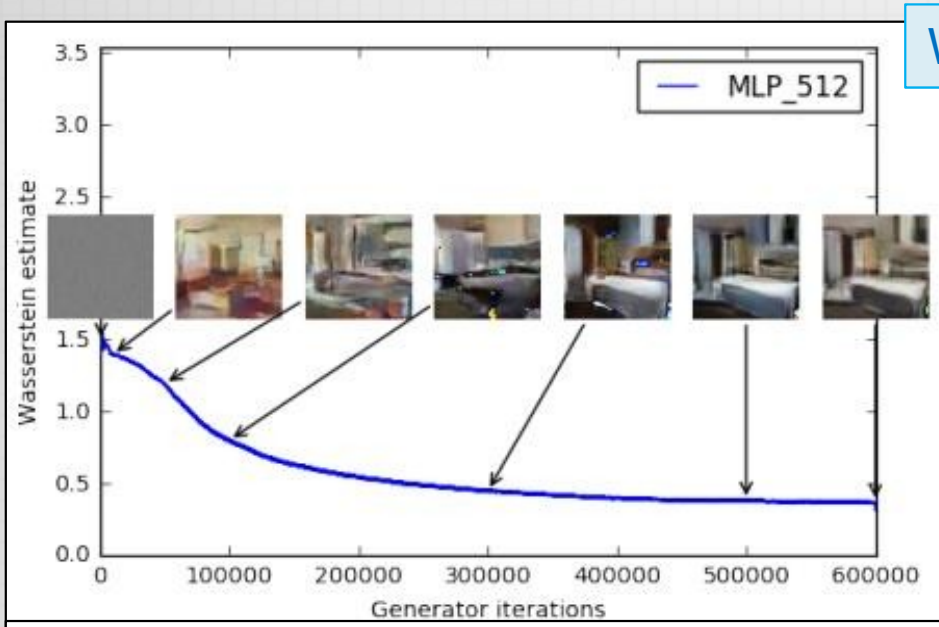
Require: : w_0, initial critic parameters. θ_0, initial generator's parameters.

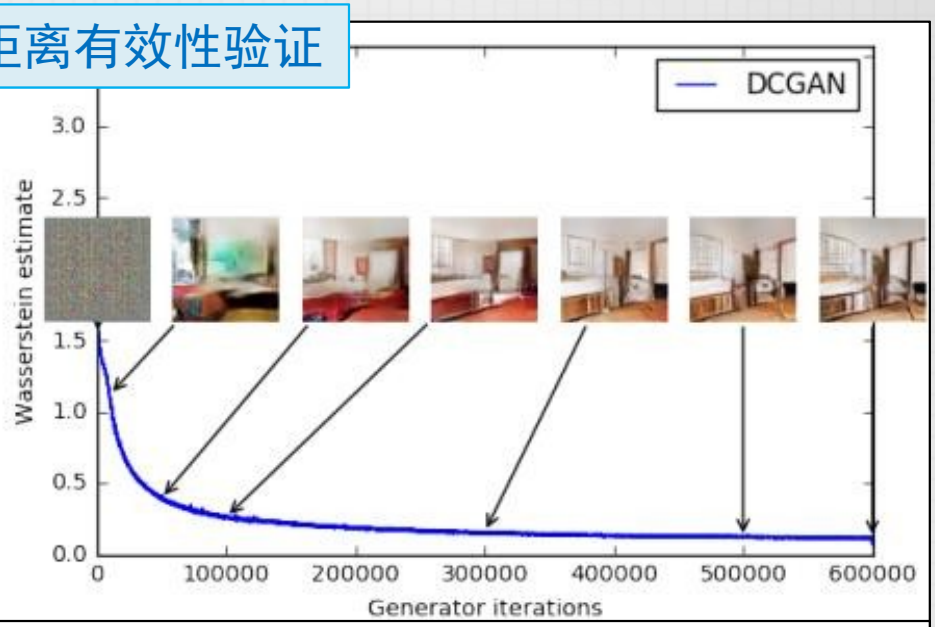
while θ has not converged **do**
  for $t = 0, \ldots, n_{\text{critic}}$ **do**
    Sample $\{x^{(i)}\}_{i=1}^{m} \sim \mathbb{P}_r$ a batch from the real data.
    Sample $\{z^{(i)}\}_{i=1}^{m} \sim p(z)$ a batch of prior samples.
    $g_w \leftarrow \nabla_w \left[\frac{1}{m}\sum_{i=1}^{m} f_w(x^{(i)})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

1. 求max·优化问题，得W距离；

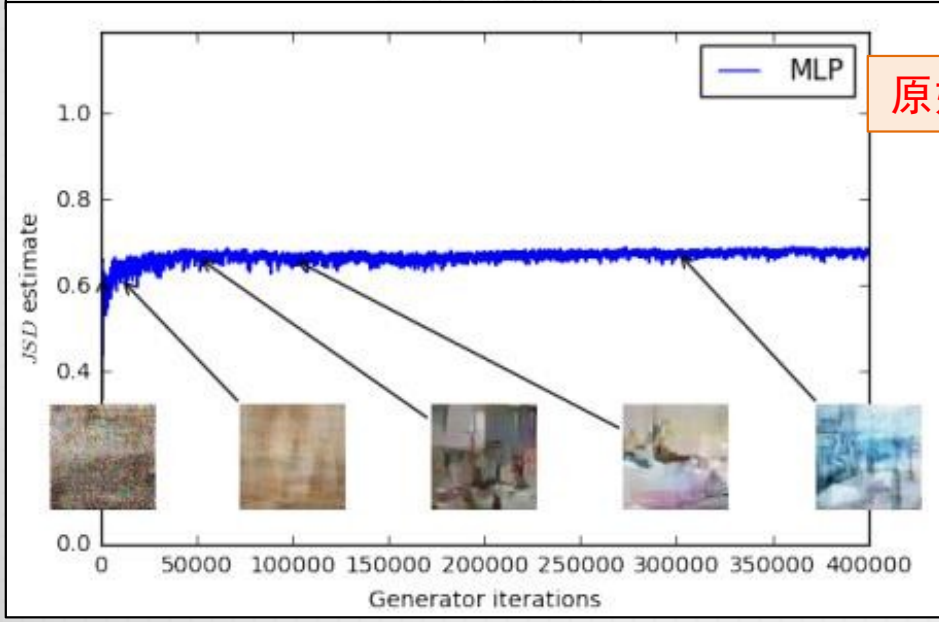
2. 求min·优化问题，最小化W距离；

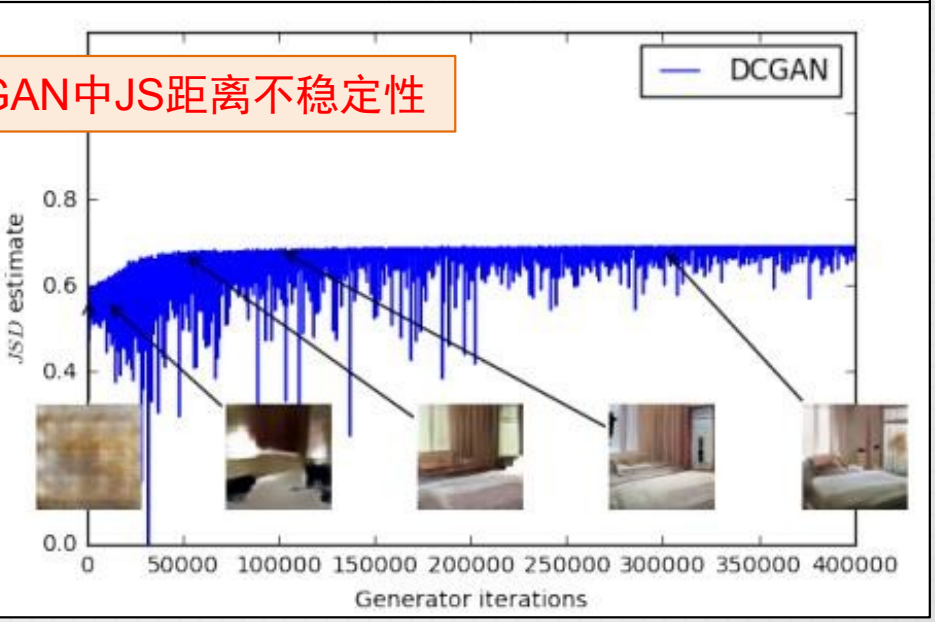
W距离有效性验证





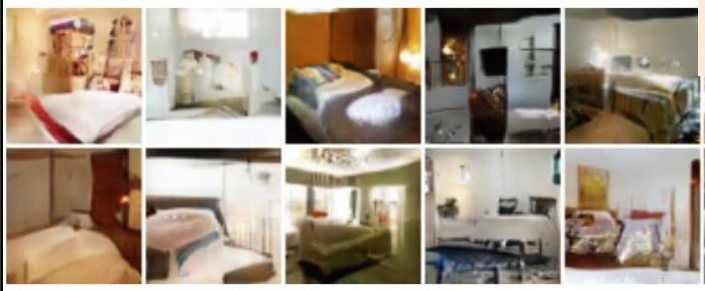
原始GAN中JS距离不稳定性



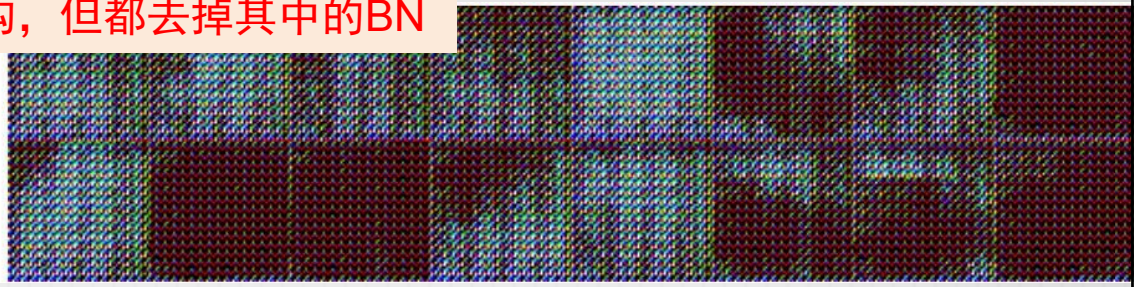


W距离不需要特殊设计网络结构的优越性

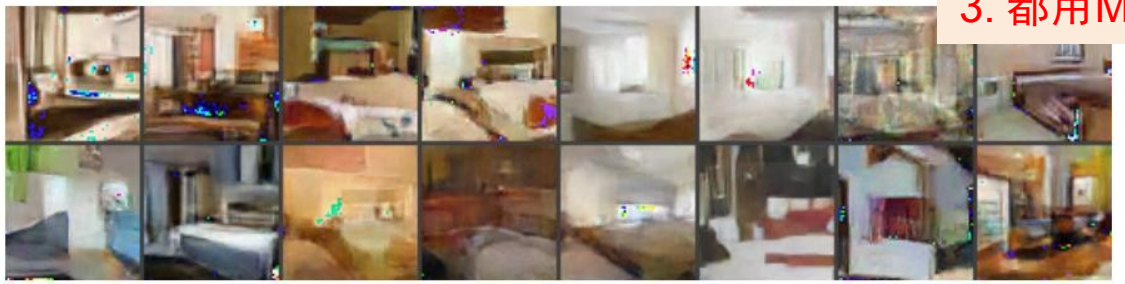
1. 都用DCGAN结构，只是距离度量不一样



2. 都用DCGAN结构，但都去掉其中的BN



3. 都用MLP结构



总结

1. KL距离、JS距离、TV距离的不合理性
2. Wasserstein距离的合理性、连续性及证明
3. Wassertein距离的近似计算与近似条件
4. WGAN的优化目标与优化过程
5. WGAN与GAN相比的优点

Thank you.