本文是关于 Wasserstein 距离在生成模型中的应用的一个总结，第一部分讲 wasserstein 距离的定义和性质，第二部分讲利用 W1 距离对偶性提出的 WGAN 和 improved WGAN，第三部分包括 ICLR18 的两篇文章，讲不依赖对偶性，可以泛化到利用 W1 距离以外的 Wasserstein 距离来产生生成模型的 WAE，以及用 NN 来模拟 wasserstein 距离的思想。

一．Wasserstein 距离

衡量两个分布的距离常用的有两种：optimal transport 以及 f-divergence（包括 kl 散度，js 散度等）。f-divergence 的定义如下，P 和 Q 是两个不同的分布，则

$$D_f(P||Q) = \int_x q(x) f\left(\frac{p(x)}{q(x)}\right) dx$$

其中 f(x)可以是任何满足 1. f is convex 2. f(1) = 0 的函数。可以证明，当 P 和 Q 完全相同，也就是说取任意的 x，都有 p(x) = q(x)，$D_f(P||Q)=0$。当 P 和 Q 有差异时，由于 f 是 convex 的：

$$D_f(P||Q) = \int_x q(x) f\left(\frac{p(x)}{q(x)}\right) dx \geq f\left(\int_x \cancel{q(x)} \frac{p(x)}{\cancel{q(x)}} dx\right)$$

后者等于 0，所以 0 是 f-divergence 的最小值。当 f 取 xlogx 时，得到了 kl 散度。

而 OT 比 f-divergence 的拓扑更弱，在生成模型中这一点非常重要，因为数据的支撑集往往是输入空间中低维流形[1]，所以真实分布和生成分布很可能没有重叠，导致 f-divergence 这种捕捉分布的概率密度比的距离会失效（p 和 q 的比值在同一个 x 点计算，而不在意 p(x1)/p(x2)的大小），从而提供不了有用的信息。OT 距离也叫 wasserstein 距离、Earth-Mover（推土机）距离。

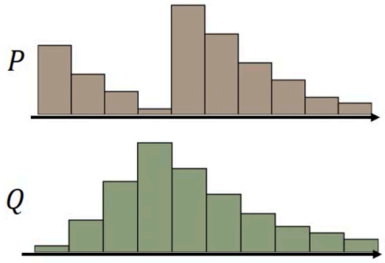
1. 定义

wasserstein 距离定义如下：

$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma \in \Pi(\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(x,y)\sim\gamma}\left[\, \|x - y\| \,\right]$$

其中$\Pi(P_r, P_g)$代表 $P_r$、$P_g$ 所有可能的联合概率分布的集合。 γ(x,y)代表了在 $P_r$ 中出现 x 同时在 $P_g$ 中出现 y 的概率，γ 的边缘分布分别为 $P_r$ 和 $P_g$。在这个联合分布下可以求得所有 x 与 y 距离的期望，存在某个联合分布使这个期望最小，这个期望的下确界（infimum）就是 $P_r$、$P_g$ 的 wasserstein 距离。

直观上看，如果两个分布是两堆土，希望把其中的一堆土移成另一堆土的位置和形状，有很多种可能的方案。

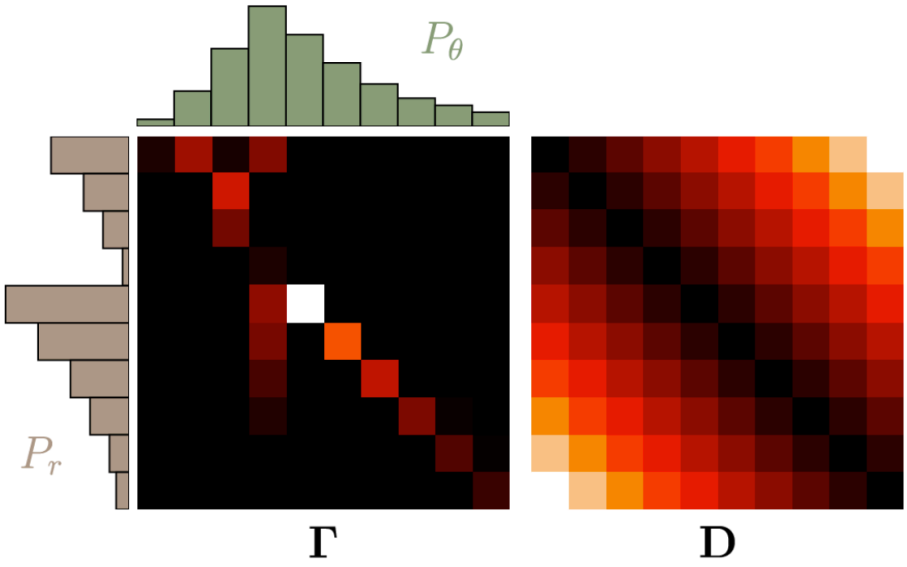



每一种方案可以对应于两个分布的一种联合概率分布，γ(x,y)代表了在 $P_r$ 中从 x 的位置移动 γ(x,y)的土量到 $P_g$ 中的 y 位置，对所有的 x 按 γ 移动，则可将分布 $P_r$ 转化成 $P_g$。推土代价被定义为移动土的量乘以土移动的距离，在所有的方

案中，存在一种推土代价最小的方案，这个代价就称为两个分布的 wasserstein 距离。设置$\mathbf{\Gamma} = \gamma(x, y)$，$\mathbf{D} = \|x - y\|$，其中$\mathbf{\Gamma}, \mathbf{D} \in \mathbb{R}^{l \times l}$，则 em 距离可以重写为：

$$\mathrm{EMD}(P_r, P_\theta) = \inf_{\gamma \in \Pi} \langle \mathbf{D}, \mathbf{\Gamma} \rangle_{\mathrm{F}}$$

其中<,>F 为内积符号。




2. Kantorovich-Rubinstein duality

当$\mathbf{D} = \|x - y\|$时，找 wasserstein 距离的问题其实是一个线性规划的问题。线性规划问题是指在线性的约束条件下找一个线性目标函数的最优化解（极大解或者极小解）。包括三个部分

1. 一个需要极大化的线性函数

$$c_1 x_1 + c_2 x_2$$

2. 以下形式的问题约束

$$a_{11} x_1 + a_{12} x_2 \leq b_1$$
$$a_{21} x_1 + a_{22} x_2 \leq b_2$$
$$a_{31} x_1 + a_{32} x_2 \leq b_3$$

3. 非负变量

$$x_1 \geq 0$$
$$x_2 \geq 0$$

在本问题中，可以将 Γ 和 D 两个矩阵展成一维：x = vec(Γ), c = vec(D)。找到 x 以最小化代价 $z = c^T x$，其中 $c \in R^n$。同时 x 需要满足约束条件 Ax = b，其中 $A \in R^{mxn}$, $b \in R^m$, $x \geq 0$。其中 $n = l^2$，$m = 2l$。

为了得到这个约束条件，令 b = $\begin{bmatrix} P_r \\ P_\theta \end{bmatrix}$。A 则需要设置为 m*n 的矩阵，挑出 x 中适当位置的值得到两个边缘分布。由于

如果像本问题中，随机变量只有一维，在这个维度上有有限个离散的状态，可以直接用解线性规划问题的方式来求解。然而在解实际问题中，比如学习图片的分布时，随机变量有上千个维度，直接计算几乎是不可能的。但是由于我们需要的只是 z 的最小值，并且利用 z 求出生成分布 $P_\theta$，而不一定需要求出 x（Γ）。所以我们可以对 z 进行关于 $P_\theta$ 的梯度下降 $\nabla_{P_\theta}\mathrm{EMD}(P_r,P_\theta)$，但是由于 $P_\theta$ 包含在优化的约束条件里，所以无法直接进行梯度下降。

由于线性规划问题都有一个对偶问题，找到本问题的对偶形式为：

| **primal form :** | | | | **dual form :** | | | |
|---|---|---|---|---|---|---|---|
| minimize | $z$ | $=$ | $\mathbf{c}^T\mathbf{x},$ | maximize | $\tilde{z}$ | $=$ | $\mathbf{b}^T\mathbf{y},$ |
| so that | $\mathbf{A}\mathbf{x}$ | $=$ | $\mathbf{b}$ | so that | $\mathbf{A}^T\mathbf{y}$ | $\leq$ | $\mathbf{c}$ |
| and | $\mathbf{x}$ | $\geq$ | $\mathbf{0}$ | | | | |

将新变量 y 作为未知变量，将最小化问题转变成了最大化问题。经过转化后，z*的取值就直接取决于 b（包含 $P_\theta$）。可以看到这个$\tilde{z}$是 z 的下限（实际上可以证明$\tilde{z}$的最大值无限接近于 z 的最小值）：

$$z = \mathbf{c}^T\mathbf{x} \geq \mathbf{y}^T\mathbf{A}\mathbf{x} = \mathbf{y}^T\mathbf{b} = \tilde{z}$$

所以目标变成了找到 y*使$\tilde{z}$* = $b^T$y*最大，$\tilde{z}$*即为两个分布的 wasserstein 距离，定义 y* = $\begin{vmatrix} f \\ g \end{vmatrix}$，则 wasserstein 距离可以表达为如下形式：

$$\mathrm{EMD}(P_r,P_\theta) = \mathbf{f}^T P_r + \mathbf{g}^T P_\theta$$

由于存在约束条件 $A^T$y ≤ c，可以得到 $f(x_i) + g(x_j) \leq D_{ij}$。由于当 i = j 时，$D_{ij}$ =0，则 $f(x_i) \leq -g(x_j)$，因为两个分布一直非负，所以要最大化$\tilde{z}$，就要最大化 $\sum_i f(xi) + g(xi)$，这个求和式在 f = -g 时取到最大值 0。

$$\begin{matrix} & \left[\begin{array}{ccc|ccc|c|ccc} \mathbf{D}_{1,1} & \mathbf{D}_{1,2} & \dots & \mathbf{D}_{2,1} & \mathbf{D}_{2,2} & \dots & \dots & \mathbf{D}_{n,1} & \mathbf{D}_{n,2} & \dots \end{array}\right] \Big\} \mathbf{c}^T \\ \mathbf{y} \left\{ \begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_n) \\ \hline g(x_1) \\ g(x_2) \\ \vdots \\ g(x_n) \end{bmatrix} \right. & \left[\begin{array}{ccc|ccc|c|ccc} 1 & 1 & \dots & 0 & 0 & \dots & \dots & 0 & 0 & \dots \\ 0 & 0 & \dots & 1 & 1 & \dots & \dots & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 0 & \dots & \dots & 1 & 1 & \dots \\ \hline 1 & 0 & \dots & 1 & 0 & \dots & \dots & 1 & 0 & \dots \\ 0 & 1 & \dots & 0 & 1 & \dots & \dots & 0 & 1 & \dots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \dots & \vdots & \vdots & \ddots \\ 0 & 0 & \dots & 0 & 0 & \dots & \dots & 0 & 0 & \dots \end{array}\right] \Big\} \mathbf{A} \end{matrix}$$

f = -g 时可得到 $f(x_i) - f(x_j) \leq D_{ij}$ 和 $f(x_i) - f(x_j) \geq -D_{ij}$，这表明 f 的倾斜程度要介于 1 和-1 之间，这个约束称为 lipschitz 连续性，对于连续分布，这个性质仍然保持:

$$\mathrm{EMD}(P_r,P_\theta) = \sup_{\|f\|_{L\leq 1}} \mathbb{E}_{x\sim P_r} f(x) - \mathbb{E}_{x\sim P_\theta} f(x)$$

而实际上，以上结论成立的前提是 1-wasserstein distance，p-wasserstein

distance 的定义如下：

$$W_p(\mu,\nu) := \inf_{\gamma\in\Gamma(\mu,\nu)}\left(\int_{\mathbb{R}^n\times\mathbb{R}^n} d(x,y)^p d\gamma(x,y)\right)^{1/p} = \left(\inf_{\xi}\mathbf{E}[d(x,y)^p]\right)^{1/p}$$

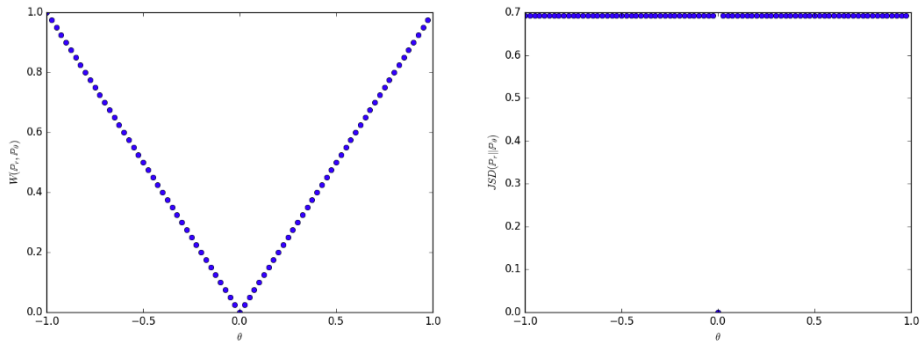
d(x,y)是 $R^l$ 上的任意距离，比如 L1 距离，欧氏距离。当 c(x,y) = d(x,y)时，上述的 Kantorovich-Rubinstein duality 成立，而当 $c(x,y) = d(x,y)^p$ 时，$f(x_i) - f(x_j) \leqslant D_{ij}$ 仍然成立，但是无法确定 f 的斜率的范围。

## 二．WGAN

WGAN 的提出是为了克服普通 GAN 出现的梯度消失的问题。wasserstein 距离与 JS 散度，KL 散度的区别在于，这个距离在两个分布不重叠的时候也是连续的。在二维空间中存在两个分布，P0 是在（0，1）上的均匀分布，P1 是在（θ,1）上的均匀分布，那么根据定义可得：

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

可以看到当 θ 趋近于 0 时，在 wasserstein 距离衡量下，P1 会趋近于 P0。而其他距离在两个分布不重叠的时候是常量，在两个分布重叠时有一个突变。所以在低维流形上学习概率分布时，可以通过 EM 距离进行梯度下降，而不能通过 JS 距离，因为它作为 loss function 是不连续的。




WGAN 使用出了上节提到的代替公式：

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

L<1 代表 f 是一个 1-Lipsschitz 函数。K-Lipsschitz 函数定义为对于 K >0，||f(x1) − f(x2)||≤K||x1 − x2||。在这里 K=1，也就是说 f 的导数不能超过 1（如果 K 取 k 的话，求得的距离则是 k 倍的 EM 距离。），限制了 f 不能改变的太快，在这个前提下，找到 f 使 $\mathbb{E}_{x\sim\mathbb{P}_r}[f(x)] - \mathbb{E}_{x\sim\mathbb{P}_\theta}[f(x)$最大，这个上确界（supremum）即为 wasserstein 距离。f(x)就可以用 nn 来拟合了，类似于普通 GAN 的判别器一样。

要限制 f 是一个 k-Lipsschitz 函数，WGAN 给出的方案是做 weight clipping，

即限制构成 f 的 NN 的参数 w 的范围在[-c,c]之间，经过梯度下降更新后，一旦某个 w 超过了这个范围，则把 w 修剪为-c 或 c。这样因为 w 的范围固定，那么随着输入的改变，输出的改变一定在一个有限的范围内，存在 k 使 f 满足 K-Lipsschitz。

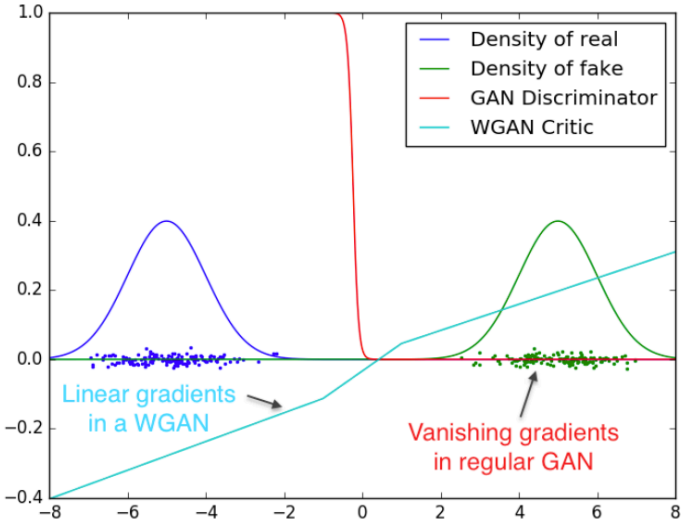
WGAN 的算法如下，与普通 GAN 对比，判别器和生成器的 loss function 都去掉了 log，判别器的最后一层去掉了 sigmoid（因为不用限制 D 的输出是 0~1），以及为判别器网络增加了 clip

**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:** : $\alpha$, the learning rate. $c$, the clipping parameter. $m$, the batch size. $n_{\text{critic}}$, the number of iterations of the critic per generator iteration.
**Require:** : $w_0$, initial critic parameters. $\theta_0$, initial generator's parameters.

**while** $\theta$ 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**

使用 WGAN 训练出来的最优判别器没有梯度消失的问题，下图是两个正态分布，可以看到 GAN 的最优判别器得到了一个 sigmoid function，完全区分开了两种分布的样本，基于本章开头的理论分析，梯度消失已经发生了，从图上也可以直观的看到，D(x)在取样处的梯度为常数，没有有效的梯度信息传给生成器；而 WGAN 限制了判别器不能太陡峭，找到了一条接近直线的线，梯度是线性的。




三．WAE 以及用 NN 学习 wasserstein 距离的方法

3.1 问题

目前的一些生成模型可以看做是用不同的方法去最小化 Px 和 Pg 的某个距离，当 Px 未知，Pg 来自神经网络时，大部分的距离是无法直接计算的。VAE 通过最小化两个分布的 KL 散度，或者等价于最大化$E_{P_X}[\log p_G(X)]$，通过最小化变分下限得到了一个可行的理论框架，GAN 则是利用分布的 JS 散度，当判别器达到最优时便能得到两个分布 js 散度。再泛化一点，使用 f-GAN 可以最小化两个分布的 f 散度。OT 距离则是另一个选择，由于 Kantorovich-Rubinstein 对偶性，可以将 OT 距离作为对抗训练的目标用在 WGAN。

从效果上说，VAE 和 GAN 各自有其缺点，VAE 模型在理论上非常优美，而在应用的时候往往会产生模糊的图片。而 GAN 生成的效果往往令人惊艳，但是由于它缺少 encoder 结构，直接从一个不包含图像信息的高斯分布去生成图像，难以训练而且经常出现 mode collapse。尽管有很多工作去组合 VAE，GAN，但是仍没有一个集 GAN 和 VAE 的优势的统一框架。

3.2 method

WAE 这个工作则是建立了 Pg 的隐变量模型，首先从隐空间一个固定的分布 Pz 中得到 Z 的采样，然后 Z 映射到一个图像 X，则 X 的分布可表示为：

$$p_G(x) := \int_{\mathcal{Z}} p_G(x|z)p_z(z)dz, \quad \forall x \in \mathcal{X}$$

其中生成模型映射$P_G(X|Z)$（也就是 decoder），可以是一个确定性的映射，对于固定的 G，将 Z 映射到 X = G(Z)；也可以是加入了随机性的 decoder。这样就可以将 OT 距离的从找在同一个空间的两个随机变量（分别服从 Px 和 Pg）的某个联合分布转变成了，对 Px 找一个条件分布$Q(Z|X)$（也就是 encoder），使边缘概率$Q_Z(Z) := \mathbb{E}_{X\sim P_X}[Q(Z|X)]$与先验概率 Pz 相同。

**定理**：对于任何映射 G：Z→X，都存在

$$\inf_{\Gamma\in\mathcal{P}(X\sim P_X, Y\sim P_G)} \mathbb{E}_{(X,Y)\sim\Gamma}\left[c(X,Y)\right] = \inf_{Q:\, Q_Z=P_Z} \mathbb{E}_{P_X}\mathbb{E}_{Q(Z|X)}\left[c(X,G(Z))\right]$$

定理 1 的证明如下：

有三个随机变量，X 是真实图像，Y 是生成的图像，Z 是隐变量。在最优传输问题中，Γ(X,Y)是联合概率分布的集合，由于 γ(X,Y)对 X 求边缘概率是 $P_x(X)$，所以 Γ(X,Y)=Γ(Y|X) $P_x(X)$，Γ(Y|X)是一个从 X 到 Y 的非确定性的映射。定理本质就是通过 Z 分解了这个映射，将它分解为 encoding 分布 Q(Z|X)和 generating 分布 $P_G(Y|Z)$。

将(Y,Z)的联合分布记为 $P_{G,Z}(Y|Z) = P_Z(Z)P_G(Y|Z)$，则：

$\mathcal{P}_{X,Y,Z}$：(X,Y,Z)联合分布的集合，满足 X ~ $P_X$，(Y,Z) ~ $P_{G,Z}$，(Y⫫ X)|Z

$\mathcal{P}_{X,Y}$：$\mathcal{P}_{X,Y,Z}$ 在(X,Y)上的边缘概率分布的集合

$\mathcal{P}_{X,Z}$：$\mathcal{P}_{X,Y,Z}$ 在(X,Z)上的边缘概率分布的集合

$\mathcal{P}(P_X,P_G)$：$P_X,P_G$ 联合概率分布的集合

根据以上定义可得到 $\mathcal{P}_{X,Y} \subseteq \mathcal{P}(P_X,P_G)$（由于对于 $\mathcal{P}_{X,Y}$ 中任意一个 X 和 Y 的联合分布，也一定属于 $\mathcal{P}(P_X,P_G)$，因为后者完全没有约束），那么就可以找到 wasserstein 距离的上界：

$$W_c(P_X,P_G) \le W_c^{\dagger}(P_X,P_G) := \inf_{P\in\mathcal{P}_{X,Y}} \mathbb{E}_{(X,Y)\sim P}\left[c(X,Y)\right]$$

引理 2：$\mathcal{P}_{X,Y} \subseteq \mathcal{P}(P_X,P_G)$，当 $P_G(Y|Z=z)$对于所有的 z 是 Dirac 函数时，$\mathcal{P}_{X,Y} = \mathcal{P}(P_X,P_G)$

如果这个引理成立的话

(1)那么 Z → Y 使用确定性的映射，就可以找到准确的 wasserstein 距离：

$$W_c(P_X,P_G) = W_c^{\dagger}(P_X,P_G)$$

$$\begin{aligned} W_c^{\dagger}(P_X,P_G) &= \inf_{P\in\mathcal{P}_{X,Y,Z}} \mathbb{E}_{(X,Y,Z)\sim P}\left[c(X,Y)\right] \\ &= \inf_{P\in\mathcal{P}_{X,Y,Z}} \mathbb{E}_{P_Z}\mathbb{E}_{X\sim P(X|Z)}\mathbb{E}_{Y\sim P(Y|Z)}\left[c(X,Y)\right] \\ &= \inf_{P\in\mathcal{P}_{X,Y,Z}} \mathbb{E}_{P_Z}\mathbb{E}_{X\sim P(X|Z)}\left[c\big(X,G(Z)\big)\right] \\ &= \inf_{P\in\mathcal{P}_{X,Z}} \mathbb{E}_{(X,Z)\sim P}\left[c\big(X,G(Z)\big)\right]. \end{aligned}$$

这样就从找 X，Y 的联合概率分布转变成了找 X，Z 的联合概率分布，而 Z 的边缘分布是固定的，P(X,Z) = P(X)P(Z|X)，那么目的则变为了找到条件分布 P(Z|X)。

(2)当 Z → Y 这个 decoder 是非确定性的，根据引理 2 我们得到的是 wasserstein 距离的上限。

推论 3：假设条件分布 $P_G(Y|Z=z)$对每个 z 有均值 G(z)，和方差$\sigma_1^2$……$\sigma_d^2$，其中 G：Z→X。取 $c(x,y) = \|x-y\|_2^2$，则

$$W_c(P_X,P_G) \le W_c^{\dagger}(P_X,P_G) = \sum_{i=1}^{d}\sigma_i^2 + \inf_{P\in\mathcal{P}(X\sim P_X, Z\sim P_Z)} \mathbb{E}_{(X,Z)\sim P}\left[\|X-G(Z)\|^2\right]$$

证明：已知

$$W_c^{\dagger}(P_X,P_G) = \inf_{P\in\mathcal{P}_{X,Y,Z}} \mathbb{E}_{P_Z}\mathbb{E}_{X\sim P(X|Z)}\mathbb{E}_{Y\sim P(Y|Z)}\left[\|X-Y\|^2\right]$$

注意到

$$\begin{aligned} &\mathbb{E}_{Y\sim P(Y|Z)}\left[\|X-Y\|^2\right] = \mathbb{E}_{Y\sim P(Y|Z)}\left[\|X-G(Z)+G(Z)-Y\|^2\right] \\ &= \|X-G(Z)\|^2 + \mathbb{E}_{Y\sim P(Y|Z)}\left[\langle X-G(Z), G(Z)-Y\rangle\right] + \mathbb{E}_{Y\sim P(Y|Z)}\|G(Z)-Y\|^2 \\ &= \|X-G(Z)\|^2 + \sum_{i=1}^{d}\sigma_i^2. \end{aligned}$$

其中第二项，Y~P(Y|Z)中，则 Y 与这个 Y 均值 G(Z)的差异对 P(Y|Z)计算期望，肯

定为 0，第三项则刚好是 Y 的方差。所以用这种方法计算出来的距离比 wasserstein 距离多一个方差。

$$\inf_{\Gamma\in\mathcal{P}(X\sim P_X, Y\sim P_G)} \mathbb{E}_{(X,Y)\sim\Gamma}\left[c(X,Y)\right] = \inf_{Q:\, Q_Z=P_Z} \mathbb{E}_{P_X}\mathbb{E}_{Q(Z|X)}\left[c(X,G(Z))\right]$$

根据定理 1，解这个问题是在 encoder $Q(Z|X)$ 上做优化，而不是在 X 和 Y 的联合分布上。在实践中，为了找到数值解，WAE 放宽了对 Qz 的约束，训练目标为：

$$D_{\mathrm{WAE}}(P_X, P_G) := \inf_{Q(Z|X)\in\mathcal{Q}} \mathbb{E}_{P_X}\mathbb{E}_{Q(Z|X)}\left[c(X,G(Z))\right] + \lambda\cdot\mathcal{D}_Z(Q_Z, P_Z)$$

其中 encoder Q 和 decoder D 都用神经网络实现，而不同 VAE，WAE 可以使用确定性的 encoder，因为不需要为每一个数据 x 在 latent space 得到一个分布，只要 Z 的边缘分布和先验接近即可。

$\mathcal{D}_Z(Q_Z, P_Z)$ 可以是一个任意的 divergence，本文尝试了两种惩罚：基于 GAN 和基于 MMD。基于 GAN 即是选择了 JS 散度来衡量分布之间的距离 $\mathcal{D}_Z(Q_Z, P_Z) = D_{\mathrm{JS}}(Q_Z, P_Z)$。

**Algorithm 1** Wasserstein Auto-Encoder with GAN-based penalty (WAE-GAN).

**Require:** Regularization coefficient $\lambda > 0$.

Initialize the parameters of the encoder $Q_\phi$, decoder $G_\theta$, and latent discriminator $D_\gamma$.

**while** $(\phi, \theta)$ not converged **do**

Sample $\{x_1, \ldots, x_n\}$ from the training set

Sample $\{z_1, \ldots, z_n\}$ from the prior $P_Z$

Sample $\tilde{z}_i$ from $Q_\phi(Z|x_i)$ for $i = 1, \ldots, n$

Update $D_\gamma$ by ascending:

$$\frac{\lambda}{n}\sum_{i=1}^{n} \log D_\gamma(z_i) + \log\left(1 - D_\gamma(\tilde{z}_i)\right)$$

Update $Q_\phi$ and $G_\theta$ by descending:

$$\frac{1}{n}\sum_{i=1}^{n} c\left(x_i, G_\theta(\tilde{z}_i)\right) - \lambda\cdot\log D_\gamma(\tilde{z}_i)$$

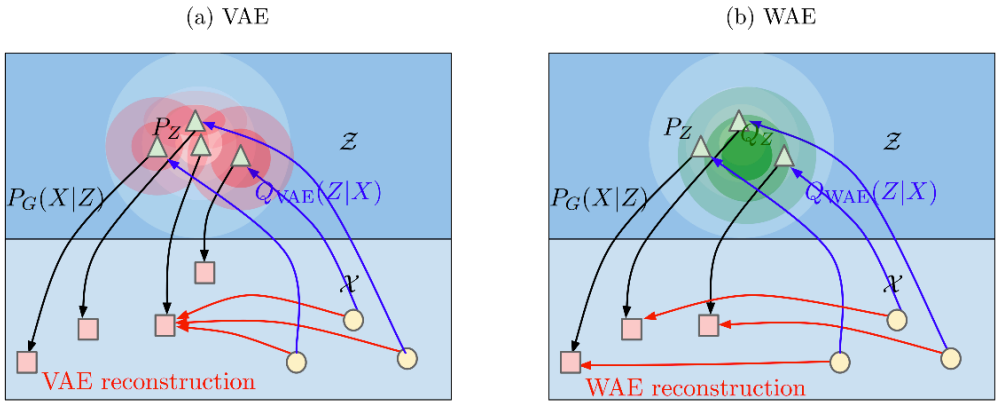
**end while**

3.3 related work

VAE：

VAE 和 WAE 都是最小化两项：重建误差、正则化项，其中第二项是通过惩罚 Pz 和通过 encoder 得到的潜在变量 z 的分布 Q 之间的距离得到的。VAE 对所

有的样本 x 强迫 Q(Z|X=x)匹配先验 Pz，如下图 a 所示，每一个红色的圆都被迫去匹配而中间白色的圆形，这样会导致红色的圆距离越来越近，相互交叉。而 WAE 是强迫生成的潜在变量 z 在 Px 分布下的期望$Q_Z := \mathbb{E}_{P_X}[Q(Z|X)]$匹配先验 Pz，如下图 b 所示的绿色的圆。这样的结果是可以不同的样本可以和其他样本保持距离。




WAE 的 decoder 的任务是精确的重构出训练样本，encoder 的任务除了上文描述的最小化生成的潜在变量和其先验分布的距离之外，还要保证潜在变量里面维持了训练样本中充足的信息以重建。

此外，WAE 可以使用确定性的 encoder，因为不需要为每一个数据 x 在 latent space 得到一个分布；VAE 只能使用高斯 encoder。

当使用$c(x,y) = \|x-y\|_2^2$时，WAE-GAN 等同于 AAE。WAE 可以看做 AAE 的泛化 1.在输入空间可以使用任意的代价函数 2. 在隐空间可以使用任意度量差异的方法，比如 MMD。

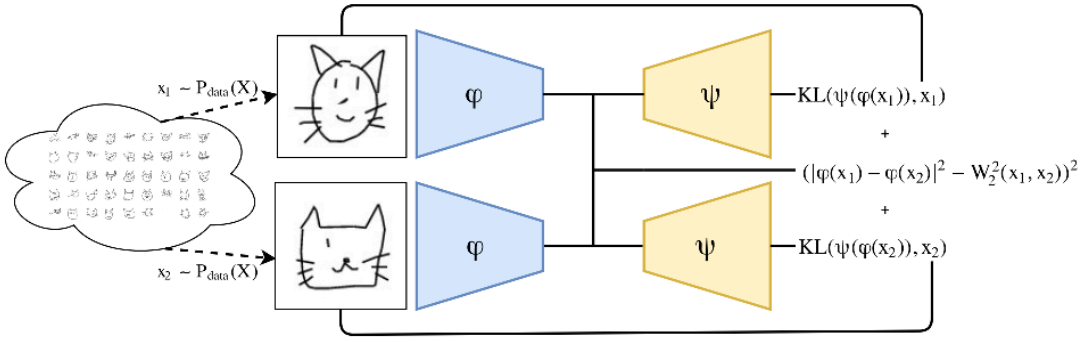
WGAN：

WGAN 最小化 1-Wasserstein distance，而不能对其他的代价 Wc 优化；同时 WGAN 不具有 encoder 结构，容易出现 model missing 的问题

同样在今年 ICLR 发表的 Deep Wasserstein Embedding，想用一种比较简单粗暴的方式直接学习 Wasserstein 距离。DWE 适用于分布已知，并且可以计算 Wasserstein 距离的情况，学习的目的是节省计算 Wasserstein 距离的时间。

DWE 就是想要用有监督的方法学习一个 embedding 的表示（同时在这个 embedding 层可以方便的进行距离计算）。需要一些预先计算好的数据集包括成对的多个直方图$\{x_i^1, x_i^2\}_{i\in 1,\dots,n}$和每对直方图对应的 wasserstein 距离$\{y_i = W_2^2(x_i^1, x_i^2)\}_{i\in 1,\dots,n}$。如果想要使用 NN 直接得到两个分布的 wasserstein 距离，一种直观的方法是将 x1 和 x2 作为输入，将距离 y 作为输出，然而这样得到的 y 不具有良好的可解释性，同时会产生不对称的距离。另一种方法则是直接对 x1 和 x2 用同一个网络进行（对称的）encode，得到含有分布信息的 embedding 结果，并在这个新的空间中模仿 wasserstein 距离学习一个欧式距离（contrastive loss）。

本文想要学习一个 embedding 网络 Φ，输入是一个直方图（即分布），将其映射到一个给定的欧几里得空间，并且希望 embedding 后能够保持 wsaaerstein 空间的几何性质。同时将 embedding 后的结果输入到 decoder 网络，计算一个重建损失，这样的做可以迫使 embedding 网络学到更多的信息来产生一个好的重建结果，使 embedding 得学习变得更简单。




DWE 方法如图所示，从一个数据分布中得到两个采样，分别作为 encoder Φ 的输入，然后通过在 embedding 层使欧氏距离的平方模仿 wasserstein 距离，用另一个 decoder 网络 Ψ 重构输入的图片，用 KL 散度计算重建误差（选择 KL 散度是因为 x 分布）。训练的目标函数为：

$$\min_{\phi,\psi} \sum_i \left| \|\phi(x_i^1) - \phi(x_i^2)\|^2 - y_i \right|^2 + \lambda \sum_i \mathrm{KL}(\psi(\phi(x_i^1)), x_i^1) + \mathrm{KL}(\psi(\phi(x_i^2)), x_i^2)$$

注意，该方法对数据集的利用不同于 WAE，如果将随机变量 X（如图片）展成长度为 d 的 vector，WAE 要学习一个 d 维随机变量的分布 P(X)（也是一般的生成模型要学习的目标），而 DWE 则是将这个 vector 作为了一个离散分布，有 d 个取值的直方图，对每一个通道 n，计算一次 wasserstein 距离。比如对于 mnist 的图片，由于是灰度图，每一个灰度值可以当做概率分布，进行 softmax 后即可归一。