

# Implicit Manifold Learning on Generative Adversarial Networks

Kry Yik Chau Lui<sup>1</sup> Yanshuai Cao<sup>1</sup> Maxime Gazeau<sup>1</sup> Kelvin Shuangjian Zhang<sup>1</sup>

## Abstract

This paper raises an implicit manifold learning perspective in Generative Adversarial Networks (GANs), by studying how the support of the learned distribution, modelled as a submanifold  $\mathcal{M}_\theta$ , perfectly match with  $\mathcal{M}_r$ , the support of the real data distribution. We show that optimizing Jensen-Shannon divergence forces  $\mathcal{M}_\theta$  to perfectly match with  $\mathcal{M}_r$ , while optimizing Wasserstein distance does not. On the other hand, by comparing the gradients of the Jensen-Shannon divergence and the Wasserstein distances ( $W_1$  and  $W_2^2$ ) in their primal forms, we conjecture that Wasserstein  $W_2^2$  may enjoy desirable properties such as reduced mode collapse. It is therefore interesting to design new distances that inherit the best from both distances.

## 1. Introduction

Unsupervised learning at present is largely about learning a probability distribution of data, either explicitly or implicitly. This is often achieved by parametrizing a probability distribution  $\mathbb{Q}_\theta$ , that is close to the real data distribution  $\mathbb{P}_r$  in some sense. The closeness criterion is typically an integral probability metric (e.g. Wasserstein distance) or an  $f$ -divergence (e.g. KL divergence). Slightly modifying Arjovsky & Bottou (2017)'s definition of *perfectly aligned* (left in figure 1), we say two manifolds  $\mathcal{M}_\theta$  and  $\mathcal{M}_r$  are *positively aligned* if the set  $\mathcal{M}_\theta \cap \mathcal{M}_r$  has a positive measure (center in figure 1).<sup>1</sup> In the context of generative modeling, two properties are desired for the closeness criterion. First, it should encourage the support of  $\mathbb{Q}_\theta$ , modelled as  $\mathcal{M}_\theta$ , to positively align with  $\mathcal{M}_r$ . This is a geometry problem, and it may be related to sample quality (more realistic generated samples). Second, it should make  $\mathbb{Q}_\theta$  and  $\mathbb{P}_r$  probabilistically similar, so samples from  $\mathbb{Q}_\theta$  reflect the multi-modal nature of  $\mathbb{P}_r$ . This is a probability problem,

<sup>1</sup>Borealis AI, Toronto, Canada. Correspondence to: Kry Yik Chau Lui <yikchau.y.lui@rbc.com>.

ICML 2017 Workshop on Implicit Models. Copyright 2017 by the author(s).

<sup>1</sup>Intuitively,  $\mathcal{M}_\theta$  and  $\mathcal{M}_r$  are the same on part of the space.

and it may be related to sample diversity (less mode dropping). The importance of the latter is well recognized (Arjovsky et al., 2017; Arora et al., 2017). The first geometric property is desired because  $\mathcal{M}_r$  might encode important constraints satisfied by real data. Consider natural images for example, samples from a learned distribution  $\mathbb{Q}_\theta$  are likely to be sharp looking if they are on  $\mathcal{M}_\theta \cap \mathcal{M}_r$ . In practice,  $\mathbb{P}_r$  is often supported on a much lower dimensional submanifold  $\mathcal{M}_r$ . For instance, the space of celebrity faces is a tiny submanifold in  $\mathbb{R}^{3 \times 64 \times 64}$  with potentially very complicated geometry. The dimensionality and geometric complexity can make the positive alignment between  $\mathcal{M}_\theta$  and  $\mathcal{M}_r$  very hard. If our goal is to generate realistic samples that respect the implicit constraints in real data, the emphasis of unsupervised learning should not only be learning the probability distribution  $\mathbb{P}_r$ , but also the manifold  $\mathcal{M}_r$ . In other words, there is an implicit manifold learning problem embedded in the explicit task of generative model learning.

Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) is a popular implicit generative model that offers great flexibility on the choice of objective functions. Extensive research (Nowozin et al., 2016; Arjovsky et al., 2017; Li et al., 2017; Bellemare et al., 2017; Berthelot et al., 2017) has been done on GANs loss function to improve training stability and mode collapse. This paper explores existing loss functions from a different perspective, namely implicit manifold learning. We show that optimizing Wasserstein distance does not guarantee positive alignment between  $\mathcal{M}_\theta$  and  $\mathcal{M}_r$ , while optimizing Jensen-Shannon divergence does. Furthermore, we attempt to clarify geometric and probabilistic properties of the Wasserstein  $W_1$ ,  $W_2^2$  metrics and Jensen-Shannon divergence, by comparing their theoretical gradients. We conjecture that  $W_2^2$  has richer geometric properties than  $W_1$ , leading to adaptive gradient update and reduced mode collapse.

## 2. Preliminaries and Definitions

Let  $\mathcal{X}$  be a compact metric space endowed with Borel  $\sigma$ -algebra  $\Sigma$ . For a probability measure  $\mu$  on  $\mathcal{X}$ , let  $\text{supp}(\mu)$  denote its support, where  $\text{supp}(\mu) := \overline{\{B \in \Sigma | \mu(B) > 0\}}$ . We work with probability distributions whose supports are  $k$ -dimensional smooth manifolds in the ambient space  $\mathbb{R}^n$ .

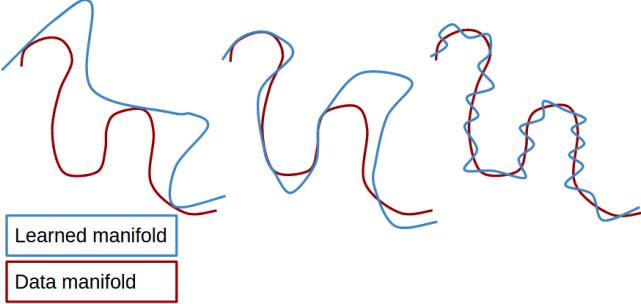


Figure 1. Concepts illustrations. Two manifolds (**left**) perfectly aligned at 3 points; (**center**) positively aligned on 3 regions (Jensen-Shannon  $\text{JSD} < \log 2$ ); (**right**) intersect transversally at many points (Wasserstein  $W_p < 0.01$ ).

Let  $\text{supp}(\mathbb{P}) = \mathcal{M}_{\mathbb{P}}^k$  and  $\text{supp}(\mathbb{Q}) = \mathcal{M}_{\mathbb{Q}}^k$  (When  $k = n$ ,  $\mathcal{M}^k = \mathbb{R}^n$ ). We focus on two probability distances in this paper, the Jensen-Shannon divergence (JSD):

$$\text{JSD}(\mathbb{P}, \mathbb{Q}) = \frac{1}{2} \text{KL}(\mathbb{P}||\mathbb{Q}_m) + \frac{1}{2} \text{KL}(\mathbb{Q}||\mathbb{Q}_m),$$

where  $\mathbb{Q}_m = \frac{1}{2}(\mathbb{P} + \mathbb{Q})$  with  $p, q$  and  $q_m$  denoting densities of  $\mathbb{P}, \mathbb{Q}$  and  $\mathbb{Q}_m$ ;

and Wasserstein  $p$ -distance  $W_p$  ( $1 \leq p < \infty$ ):

$$W_p(\mathbb{P}, \mathbb{Q}) = \left( \inf_{\gamma \in \Pi(\mathbb{P}, \mathbb{Q})} \int \|x - y\|^p d\gamma(x, y) \right)^{\frac{1}{p}} \quad (1)$$

where  $\Pi(\mathbb{P}, \mathbb{Q})$  denotes the collection of all probability measures on  $\mathcal{M}_{\mathbb{P}}^k \times \mathcal{M}_{\mathbb{Q}}^k$  with marginals  $\mathbb{P}$  and  $\mathbb{Q}$  on the first and second variables respectively. As  $\mathcal{M}_{\mathbb{P}}^k$  and  $\mathcal{M}_{\mathbb{Q}}^k$  have the same dimensions, we simplify their notations as  $\mathcal{M}_{\mathbb{P}}$  and  $\mathcal{M}_{\mathbb{Q}}$  when contexts are clear. Monge (Monge, 1781) originally formulated the distance as:<sup>2</sup>

$$W_p(\mathbb{P}, \mathbb{Q}) = \inf_{T_*(\mathbb{P}) = \mathbb{Q}} [\mathbb{E}_{x \sim \mathbb{P}_r} \|x - T(x)\|^p]^{1/p} \quad (2)$$

where  $T_*(\mathbb{P}) = \mathbb{Q}$  means a Borel map  $T$  pushes forward  $\mathbb{P}$  to  $\mathbb{Q}$ , i.e.  $\int_{T^{-1}(B)} p = \int_B q$  for any Borel set  $B \subset \mathcal{M}^k$ . Note the infimum in equation (2) is taken over the space of Borel maps while in equation (1) the infimum is searched over the space of probability measures. We consider the cases whenever the infimum is achieved by an optimal transport map  $T_p$ . For example when  $p = 2$ , for each  $\mathbb{Q}$ , by Brenier's theorem (McCann, 2001; McCann & Guillen, 2011) there exists an optimal transport map  $T_2$  such that  $W_2^2(\mathbb{P}, \mathbb{Q}) = \mathbb{E}_{x \sim \mathbb{P}_r} [\|x - T_2(x; \mathbb{Q})\|^2]$ .

### 3. Sample Quality

Since its introduction, sample quality in Generative Adversarial Nets (GANs) has improved dramatically (Goodfellow et al., 2014; Radford et al., 2015; Berthelot et al., 2017), and it arguably generates the most realistic looking

<sup>2</sup> Historically, Monge formulated  $W_1$  only.

images nowadays. However, little theory exists to explain why this is the case (Goodfellow, 2016). One reason is a precise definition of “sharp looking” is missing.

When  $\mathbb{P}_r$  is the distribution of natural images, its support  $\text{supp}(\mathbb{P}_r)$  is probably sufficiently structured that it can be modeled by a  $k$ -dimensional submanifold  $\mathcal{M}_r$  in the ambient space  $\mathbb{R}^n$  (Narayanan & Mitter, 2010). Now pick a sample  $x$  from  $\mathcal{M}_r$  and consider its perturbation,  $\tilde{x} = x + \epsilon$ , where  $\epsilon \in \mathbb{R}^n$  and  $\|\epsilon\|$  fixed. Depending on  $\epsilon$ 's direction, some  $\tilde{x}$  might look realistic while others may not. When  $\|\epsilon\|$  increases, the difference becomes more vivid. This is remarkably similar to the fact that some  $\tilde{x}$  travel along  $\mathcal{T}_x \mathcal{M}_r$  the tangent space of  $\mathcal{M}_r$  at  $x$  while others go off  $\mathcal{M}_r$ . When it is on  $\mathcal{T}_x \mathcal{M}_r$ ,  $\tilde{x}$  looks sharper. When it goes off,  $\tilde{x}$  no longer looks natural. This motivates:

**Definition 3.1** (Realistic Samples). We say  $\mathbb{Q}$  generates realistic  $\mathbb{P}_r$  samples if  $\mathcal{M}_q = \text{supp}(\mathbb{Q})$  positively aligns with  $\mathcal{M}_r = \text{supp}(\mathbb{P}_r)$ . In other words, samples from  $\mathbb{Q}$  are realistic with respect to  $\mathbb{P}_r$  if they lie exactly on  $\mathcal{M}_r$ .

In GANs,  $\mathbb{Q}_\theta$  is the distribution implicitly parametrized by the generator  $G_\theta$ . Ideally,  $\mathbb{Q}_\theta$  can generate indistinguishable samples from  $\mathbb{P}_r$  after training. We next show optimizing JSD successfully will necessarily positively align  $\mathcal{M}_\theta$  and  $\mathcal{M}_r$ , hence  $\mathbb{Q}_\theta$  can generate at least some realistic samples. This is intuitive, since whenever  $\mathcal{M}_r$  and  $\mathcal{M}_\theta$  do not positively align, JSD is maxed out. We assume the following to translate our intuitions to theorems:

**Assumption A:**  $\mathbb{P}_r$  and  $\mathbb{Q}$  are compactly supported on  $\mathcal{M}_r^k$  and  $\mathcal{M}_q^k$ ,  $k < n$ , satisfying  $\mathcal{L}^k(\mathcal{M}_r^k), \mathcal{L}^k(\mathcal{M}_q^k) > 0$ .<sup>3</sup>

**Assumption B:**  $\mathbb{P}_r$  and  $\mathbb{Q}$  are absolutely continuous with respect to  $\mathcal{L}^k(\mathcal{M}_r^k)$  and  $\mathcal{L}^k(\mathcal{M}_q^k)$ , i.e., for any set  $B \subset \mathbb{R}^n$ ,  $\mathbb{P}_r(B) = \mathbb{Q}(B) = 0$  whenever  $\mathcal{L}^k(B) = 0$ .

**Definition 3.2** (Minimal common support). Under **Assumption A** and **B**, let  $0 \leq \alpha \leq \log 2$  be given. Consider the set of distributions  $\mathbb{Q}$  that achieve at most  $\alpha$  level JSD:  $\Omega^\alpha = \{\mathbb{Q} : \text{JSD}(\mathbb{P}_r, \mathbb{Q}) \leq \alpha\}$ . For any fixed  $\mathbb{P}_r$ , we define the **minimal common support** to achieve at most  $\alpha$  level JSD to be:  $\text{MCS}^\alpha(\mathbb{P}_r) = \inf_{\mathbb{Q} \in \Omega^\alpha} \mathcal{L}^k(\text{supp}(\mathbb{P}_r) \cap \text{supp}(\mathbb{Q}))$ .

When  $\mathbb{Q}$  is implicitly parametrized by neural networks with parameters  $\theta$ , the notations  $\Omega_\theta^\alpha$  and  $\text{MCS}_\theta^\alpha(\mathbb{P}_r)$  reflect their dependency on  $\theta$ . Definition 3.2 captures the worst case scenario: when  $\text{JSD} < \log 2$ , is  $\text{MCS}^\alpha(\mathbb{P}_r) > 0$ ? In other words, whenever JSD is not maxed out, can we expect  $\mathbb{Q}$  to generate some  $\mathbb{P}_r$  realistic samples with nonzero probability? The next proposition gives a positive answer.

**Theorem 3.1.** Let **Assumption A** and **B** hold and  $p_r$ , the

<sup>3</sup>  $\mathcal{L}^k$  denotes Lebesgue measure on  $\mathbb{R}^k$ . Strictly speaking,  $\mathcal{L}^k$  should be replaced by Hausdorff measure  $\mathcal{H}^k$ . When  $k = 2$ ,  $\mathcal{H}^2$  is the measure theoretic surface area.

density of  $\mathbb{P}_r$ , be bounded, then for  $\alpha \in [0, \log 2]$ ,  $\text{MCS}^\alpha > 0$ ; when  $\alpha = \log 2$ ,  $\text{MCS}^\alpha = 0$ .

Theorem 3.1 ensures  $\text{MCS}^\alpha(\mathbb{P}_r)$  is well-defined. The next corollary suggests JSD is a sensible objective to optimize when it comes to generating realistic samples.

**Corollary 3.1.** Under the assumptions in Proposition 3.1,  $\text{MCS}^\alpha(\mathbb{P}_r)$  is non-increasing with respect to  $\alpha$  on the interval  $[0, \log 2]$ .

The next theorem states optimizing Wasserstein distances does not force positive alignment. In other words, there is no guarantee that  $\mathcal{M}_r$  and  $\mathcal{M}_q$  positively align unless  $W_p(\mathbb{P}_r, \mathbb{Q}) = 0$ . This is because we can find many distributions  $\mathbb{Q}$  such that  $W_p(\mathbb{P}_r, \mathbb{Q}) < \epsilon$  but  $\mathcal{M}_r$  and  $\mathcal{M}_q$  do not positively align, however small  $\epsilon > 0$  gets. For pictorial illustrations and comparison of theorems 3.1 and 3.2, see (center) and (right) in figure 1.

**Theorem 3.2.** Let  $\epsilon > 0$  and  $\mathbb{P}_r$  be a fixed distribution. Let  $\Gamma = \{\mathbb{Q} : W_p(\mathbb{P}_r, \mathbb{Q}) < \epsilon\}$ , and consider the decomposition:  $\Gamma = \Gamma_1 \cup \Gamma_2$ , where  $\Gamma_1 = \{\mathbb{Q} : W_p(\mathbb{P}_r, \mathbb{Q}) < \epsilon; \mathcal{L}^k(\text{supp}(\mathbb{P}_r) \cap \text{supp}(\mathbb{Q})) > 0\}$  and  $\Gamma_2 = \Gamma - \Gamma_1$ . Then under Assumption A,  $\Gamma_2$  is dense in  $\Gamma$ .

As a result, the problem that  $W_p$  GANs do not necessarily generate realistic samples cannot be solved by increasing model capacity.

## 4. Sample Diversity and Adaptive Gradient

Under finite capacity, (Arora et al., 2017) shows there are mode collapse scenarios that few current training objectives in GANs can prevent. In the follow-up empirical analysis, (Arora & Zhang, 2017) raises the open problem on redesigning GANs objective so as to avoid mode collapses. A less ambitious quest is to compare the existing loss functions and identify properties related to mode dropping. Hopefully this suggests new designs that combat mode collapses. A natural place to start the comparison is with the gradients of the generator loss functions.

### 4.1. The Wasserstein $W_1$ and $W_2$ distance

There are empirical evidences showing that Wasserstein  $W_1$  GANs (Arjovsky et al., 2017) exhibit less mode collapse than Jensen-Shannon GANs. This is probably due to its geometric properties. We attempt to examine this by computing  $\nabla_\theta W_1(\mathbb{P}_r, \mathbb{Q}_\theta)$  in its primal form. If the geometric properties of  $W_1$  makes it more robust to mode dropping, then it is also interesting to investigate  $W_2$  which better reflects geometric features (Villani, 2008). While it is unclear how to apply  $W_2$  to GANs training due to its more complex dual formulation, it is instructive to analyze its theoretical gradient  $\nabla_\theta W_2^2(\mathbb{P}_r, \mathbb{Q}_\theta)$ .

**Proposition 4.1.** Let  $\mathbb{P}_r$  and  $\mathbb{Q}_\theta$  be two distributions with absolutely continuous densities on  $\mathcal{M}_r^k$  and  $\mathcal{M}_\theta^k$  in the ambient space  $\mathbb{R}^n$ , with  $k \leq n$ . We have:

$$\nabla_\theta W_2^2(\mathbb{P}_r, \mathbb{Q}_\theta) = -2 \int (x - T_2(x; \theta)) \nabla_\theta T_2(x; \theta) p_r(x) dx \quad (3)$$

Similarly, we have the following for  $W_1(\mathbb{P}_r, \mathbb{Q}_\theta)$ :

$$\nabla_\theta W_1(\mathbb{P}_r, \mathbb{Q}_\theta) = \int \pm \mathbf{I} \nabla_\theta T_1(x; \theta) p_r(x) dx \quad (4)$$

whenever both sides are well defined.  $\pm \mathbf{I}$  is a vector valued functions with codomain  $[\pm 1, \dots, \pm 1]$  where the sign depends on whether  $(x - T_1(x; \theta))_i$  is positive or negative, for  $1 \leq i \leq n$ .

Let us consider the update equation (3) with one sample point:  $\theta_{t+1} = \theta_t + 2(x - T_2(x; \theta)) \nabla_\theta T_2(x; \theta)$ . The first term  $x - T_2(x; \theta)$  gives  $W_2^2$  its geometric properties. When  $\mathcal{M}_r$  and  $\mathcal{M}_\theta$  are far away,  $\|x - T_2(x; \theta)\|$  is very big. This should strongly attracts  $\mathcal{M}_\theta$  to  $\mathcal{M}_r$  in  $\mathbb{R}^n$ . When  $\mathcal{M}_r$  and  $\mathcal{M}_\theta$  become closer,  $\|x - T_2(x; \theta)\|$  is smaller. This resembles  $L^2$  optimization in general, where the loss function offers an adaptive gradient. The third term  $p_r(x)$  provides a multi-modal weighting. The higher  $p_r(x)$ , the stronger contribution it gives to  $\nabla_\theta W_2^2(\mathbb{P}_r, \mathbb{Q}_\theta)$ . Therefore  $\mathbb{P}_r$ 's modes will drive the gradient update.

On the other hand, equation (4) for Wasserstein  $W_1$  is closer to  $L^1$  geometry. While it has the same probabilistic weighting as  $W_2^2$ , its geometric part is plainer: the first term is a signed vector  $\pm \mathbf{1}$  that does not adapt according to  $\|x - T_1(x; \theta)\|$  (how far away  $\mathcal{M}_r$  and  $\mathcal{M}_\theta$  are). However, our analysis is limited because the optimal transport maps  $T_1$  and  $T_2$  are implicitly defined. It is possible that  $\nabla_\theta T_1(q_\theta)$  and  $\nabla_\theta T_2(q_\theta)$  can cancel the above desired geometric and probabilistic properties. Nonetheless, we believe the above calculations partially clarify some of the geometric and probabilistic advantages of  $W_1$  and  $W_2^2$ .

### 4.2. The Jensen-Shannon Divergence

In light of previous section, we perform similar calculations for JSD and the reversed  $-\log D$  trick. The following assumption is needed to insure KL divergence is finite:

**Assumption C:** Let  $\mathbb{P}_r$  and  $\mathbb{Q}_\theta$  be absolutely continuous with respect to  $\mathcal{L}^n$  with equal support and  $\mathcal{L}^n(\text{supp}(\mathbb{P}_r)) > 0$ .<sup>4</sup>

**Proposition 4.2.** Let  $D^*(x) = \frac{p_r(x)}{q_{\theta_0}(x) + p_r(x)}$  be the optimal discriminator, for  $\theta_0$  fixed. Under Assumption C, we have:

$$\begin{aligned} & \nabla_\theta \mathbb{E}_{z \sim p(z)} [-\log D^*(g_\theta(z))] |_{\theta=\theta_0} \\ &= \mathbb{E}_{\mathbb{Q}_\theta} \left[ \nabla_\theta \log(q_\theta) \left( 1 + \log \left( \frac{q_m}{p_r} \right) \right) \right] \Big|_{\theta=\theta_0}, \end{aligned} \quad (5)$$

<sup>4</sup>These assumptions make sense when we convolve  $\mathbb{P}_r$  and  $\mathbb{Q}_\theta$  with an  $n$ -dimension Gaussian, as in (Arjovsky & Bottou, 2017).

and for the standard JSD:

$$\begin{aligned} & \nabla_{\theta} \mathbb{E}_{z \sim p(z)} [\log (1 - D^*(g_{\theta}(z))] |_{\theta=\theta_0} \\ &= \mathbb{E}_{Q_{\theta}} \left[ \nabla_{\theta} \log(q_{\theta}) \log \left( \frac{q_{\theta}}{q_m} \right) \right] \Big|_{\theta=\theta_0}. \end{aligned} \quad (6)$$

Like in section 4.1, we study the influence of each objective on mode collapse. We analyze equations (5) and (6) where  $q_{\theta}$  is very small and  $p_r$  is comparably large, which is often the case in early training.

First we note the influence of  $p_r(x)$  is not as obvious as in (3) or (4), as the weight factors  $|1 + \log \frac{q_m}{p_r}|$  in (5) and  $|\log \frac{q_{\theta}}{q_m}|$  in (6) involve  $q_{\theta}$  as well. Assume  $q_{\theta}$  is fixed. For equation (5) ( $-\log D$  trick), the weight factor  $|1 + \log \frac{q_m}{p_r}|$  strictly decreases as  $p_r(x)$  gets larger. This is undesired because  $p_r(x)$ 's higher probability regions contribute less to  $\nabla_{\theta} \log(q_{\theta})$ . What's worse, the regions where  $p_r(x)$  is small gets a stronger gradient. Thus, if  $q_{\theta}$  misses some modes in the first place, it may be less likely to learn those modes in later updates. In contrast, for equation (6) (standard Jensen-Shannon GAN), the weight  $|\log \frac{q_{\theta}}{q_m}|$  has the right monotonic relation: it assigns more weights to regions where  $p_r(x)$  is bigger. This suggests when  $D = D^*$ , the classical  $\nabla_{\theta}$  JSD( $p_r, q_{\theta}$ ) is better suited to look for missing modes when the gradient  $\nabla_{\theta} q_{\theta}$  does not vanish.<sup>5</sup><sup>6</sup> Similar to section 4.1, our analysis is non-conclusive because  $\nabla_{\theta} \log(q_{\theta})$ , like  $\nabla_{\theta} T_2(x; \theta)$ , is implicitly defined.

## 5. Discussions and Future Work

This paper suggests Wasserstein distances and Jensen-Shannon divergences can complement each other on two important aspects of GANs training, namely sample quality (sharpness) and sample diversity (mode collapse). Geometric property of Wasserstein distance comes from the distance between the samples  $\|x - y\|_{x \sim p_r, y \sim q_{\theta}}$ , while Jensen-Shannon divergence acts purely on the densities. Its sharpness property is due to the logarithmic weights on the densities, i.e.  $\log p_r - \log \frac{1}{2}(p_r + q_{\theta})$ , which heavily penalizes the non-positively aligned supports. To preserve both desired properties, we can either combine these two measures, say by proportional control as in (Berthelot et al., 2017) or design a new distance that operates on both samples and the probability densities.

As the empirical sample quality in Jensen-Shannon GANs

<sup>5</sup>In our preliminary experiments, when Lipschitz constraints (Gulrajani et al., 2017) is applied to standard JSD GANs,  $\nabla_{\theta} G_{\theta}(z)$  does not vanish and it trains as well as the  $-\log(D)$  trick. This is probably due to the preactivation in logit does not lie in the saturation region due to the global Lipschitz constant.

<sup>6</sup>Note this does not necessarily contradict (Arjovsky & Bottou, 2017)'s observation that  $\nabla_{\theta}$  JSD( $\mathbb{P}_r, Q_{\theta}$ ) suffers from vanishing gradient. Even if  $|\log(\frac{q_{\theta}}{q_m})| \rightarrow \infty$ , so long as  $\nabla_{\theta} q_{\theta} \rightarrow 0$  faster, we still have vanishing gradient.

does not match our theory, identifying the reasons is interesting. First, a lower bound of Jensen-Shannon divergence is optimized (Nowozin et al., 2016) in practice, instead of the divergence itself. Second, (Arora et al., 2017) points out the importance of finite sample and finite capacity when we reason GANs training. We believe a similar principle applies here. Using their definition:

**Definition 5.1** ( $\mathcal{F}$ -distance). Let  $\mathcal{F}$  be a class of functions from  $\mathbb{R}^n$  to  $[0, 1]$ . Then  $\mathcal{F}$ -distance is:

$$\begin{aligned} d_{\mathcal{F}, \log}(\mathbb{P}, \mathbb{Q}) &= \sup_{D \in \mathcal{F}} |\mathbb{E}_{x \sim \mathbb{P}}[\log(D(x))] \\ &\quad - \mathbb{E}_{x \sim \mathbb{Q}}[\log(1 - D(x))] - 2 \log(1/2). \end{aligned}$$

When  $\mathcal{F} = \{ \text{all functions from } \mathbb{R}^n \text{ to } [0, 1] \}$ ,  $d_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = \text{JSD}(\mathbb{P}, \mathbb{Q})$ . When  $\mathcal{F}$  is restricted to a set of neural nets with finite parameters, we let  $\widehat{\text{JSD}}$  denote the corresponding neural net distance. It is then natural to define a finite capacity version of definition 3.2:

**Definition 5.2** (Finite Capacity Minimal common support). Let  $0 \leq \alpha \leq \log 2$  be given. Consider the set of implicitly parametrized distributions  $Q_{\theta}$  that achieve at most  $\alpha$  level JSD:  $\Omega_{\theta}^{\alpha} = \{Q_{\theta} : \widehat{\text{JSD}}(\mathbb{P}_r, Q_{\theta}) \leq \alpha\}$ . For any fixed  $\mathbb{P}_r$ , we define the finite capacity minimal common support to achieve at most  $\alpha$  level  $\widehat{\text{JSD}}$  divergence to be:  $FMC_{\theta}^{\alpha}(\mathbb{P}_r) = \inf_{Q_{\theta} \in \Omega_{\theta}^{\alpha}} \mathcal{L}^k(\text{supp}(\mathbb{P}_r) \cap \text{supp}(Q_{\theta}))$ .

Under finite capacity and finite sample, is it important to understand if a similar conclusion like theorem 3.1 still holds. Let  $\widehat{\mathbb{P}}_r$  and  $\widehat{\mathbb{Q}}_{\theta}$  be the corresponding empirical distributions. Let  $\widehat{\alpha}_1$  and  $\widehat{\alpha}_2$  be the corresponding  $\widehat{\text{JSD}}$  values computed on finite samples<sup>7</sup>. Is it true for sufficiently regular  $\mathcal{M}_r$  and a moderately sized sample from  $\mathbb{P}_r$ :  $\widehat{\alpha}_2 < \widehat{\alpha}_1 \Rightarrow FMC_{\theta}^{\widehat{\alpha}_2}(\mathbb{P}_r) \geq FMC_{\theta}^{\widehat{\alpha}_1}(\mathbb{P}_r)$  with high probability?<sup>8</sup><sup>9</sup> More generally, what kind of neural net distance can give the above properties? Recently, (Berthelot et al., 2017) demonstrated impressive sample quality. How do their approaches positively align  $\mathcal{M}_{\theta}$  with  $\mathcal{M}_r$ ?

Moreover, since  $\mathcal{M}_{\theta}$  is parametrized by the generator, we may regularize  $G_{\theta}$  based on  $\mathcal{M}_r$ 's geometric structure. So the cost functions will include a geometric loss and a probability distance.

While we discussed implicit manifold learning under GANs framework in this paper, it is also interesting to explore this perspective with other generative models such as Variational Autoencoder (Kingma & Welling, 2013).

<sup>7</sup>  $\widehat{\alpha}_1 = \widehat{\text{JSD}}(\widehat{\mathbb{P}}_r, \widehat{\mathbb{Q}}_{\theta_1})$  and  $\widehat{\alpha}_2 = \widehat{\text{JSD}}(\widehat{\mathbb{P}}_r, \widehat{\mathbb{Q}}_{\theta_2})$ , for samples from  $\mathbb{Q}_{\theta_1}$  and  $\mathbb{Q}_{\theta_2}$

<sup>8</sup> The probability is over  $\widehat{\mathbb{Q}}_{\theta}$ ; we repeatedly sample from  $\mathbb{Q}_{\theta}$ .

<sup>9</sup> In practice, a sufficiently well trained discriminator  $D$  is used to approximate the true neural net distance.

## Supplementary Materials

### 6. Proofs

**Proposition 6.1** (Proposition 3.1 in main paper). *Let Assumption A and B hold and  $p_r$ , the density of  $\mathbb{P}_r$  be bounded, then for  $\alpha \in [0, \log 2)$ ,  $\text{MCS}^\alpha > 0$ ; when  $\alpha = \log 2$ ,  $\text{MCS}^\alpha = 0$ .*

*Proof.* The proof is divided into two parts. In the first part, we show that the minimum common support between  $\mathbb{P}_r$  and  $\mathbb{Q}_\theta$  is strictly positive for all  $\alpha \in [0, \log 2)$ . In the second part, we show the minimum common support is equal to zero for  $\alpha = \log 2$ .

Let us prove the first part by contradiction and assume that there exists an  $\alpha_0 \in [0, \log 2)$ , such that  $\text{MCS}_{\alpha_0} = 0$ . By definition of the infimum, there exists a minimizing sequence of distributions in  $\Omega_{\alpha_0}$ , denoted as  $\{\mathbb{Q}_\theta^m\}_{m=1}^\infty$ , such that  $\mathcal{L}^k(\text{supp}(\mathbb{P}_r) \cap \text{supp}(\mathbb{Q}_\theta^m)) \rightarrow 0$ , as  $m \rightarrow \infty$ . Then, by definition of the set  $\Omega_{\alpha_0}$ ,  $\text{JSD}(\mathbb{P}_r, \mathbb{Q}_\theta^m) \leq \alpha_0 < \log 2$  and there is an overlap between  $\mathbb{P}_r$  and  $\mathbb{Q}_\theta^m$ .

Without loss of generality, we assume  $\text{supp}(\mathbb{Q}_\theta^m) \subset \text{supp}(\mathbb{P}_r)$ . We define the set  $S_m := \text{supp}(\mathbb{Q}_\theta^m)$  and its complementary  $S_m^c = \text{supp}(\mathbb{P}_r) \setminus \text{supp}(\mathbb{Q}_\theta^m)$ . Moreover for each  $J > 0$ , we define  $S_m^J = \{x \in S_m : q_\theta^m(x) \leq J\}$ .

We write  $2 \text{JSD}(\mathbb{P}_r, \mathbb{Q}_\theta^m) = \sum_{j=1}^5 J_j^m$ , where the five terms are given by

$$\begin{aligned} J_1^m &:= \int_{S_k} p_r(x) \log(2p_r(x)) dx, \\ J_2^m &:= - \int_{S_k} p_r(x) \log(p_r(x) + q_\theta^k(x)) dx, \\ J_3^m &:= \text{KL}(\mathbb{Q}_\theta^m, (\mathbb{Q}_\theta^m + \mathbb{P}_r)/2) \Big|_{S_m^J}, \\ J_4^m &:= \text{KL}(\mathbb{Q}_\theta^m, (\mathbb{Q}_\theta^m + \mathbb{P}_r)/2) \Big|_{S_m \setminus S_m^J}, \\ J_5^m &:= 2 \text{JSD}(\mathbb{P}_r, \mathbb{Q}_\theta^m) \Big|_{S_m^c}, \end{aligned}$$

From the inequality  $x \log(2x) \geq -1$  for all  $x \geq 0$ , we deduce  $J_1^m \geq -\mathcal{L}^n(S_m)$ . From the boundedness of  $p_r$  by  $N$  and using the Jensen inequality applied to the convex function  $x \log(x)$ , we have  $J_2^m \geq -N\mathcal{L}^k(S_m) \log(N + 1/\mathcal{L}^k(S_m))$ . On the set  $S_m^J$ ,  $q_\theta^k \leq J$ . Therefore from the Jensen inequality, we get  $J_3^m \geq (J + N)\mathcal{L}^k(S_m^J) \cdot \min_{x \geq 0}(x \log x)/2$ . By a diagonal extraction argument, we can extract a subsequence  $\{q_\theta^{m_i}\}_{i=1}^\infty$  such that  $\mathcal{L}^k(S_{m_i}^i) \leq \frac{1}{i^2}$  and  $J_4^m \geq \log(\frac{2i}{i+N})[1 - i\mathcal{L}^k(S_{m_i}^i)] \geq \log(\frac{2i}{i+N})(1 - \frac{1}{i})$ . Finally on  $S_k^c$ ,  $q_\theta^m = 0$  and as a consequence  $J_5^m = \int_{S_m^c} p_r(x) \log(2) dx = \log(2)(1 - \mathbb{P}_r(S_m))$ .

Gathering the above inequalities, we deduce  $2 \log 2 \geq \lim_{i \rightarrow \infty} 2 \text{JSD}(\mathbb{P}_r, \mathbb{Q}_\theta^{m_i}) = \lim_{i \rightarrow \infty} J_1^{m_i} + J_2^{m_i} + J_3^{m_i} + J_4^{m_i} + J_5^{m_i} \geq 0 + 0 + 0 + \log 2 + \log 2 = 2 \log 2$ , as  $m_i \rightarrow \infty$ . Thus we deduce from the squeeze theorem that there exists a subsequence such that  $\lim_{i \rightarrow \infty} \text{JSD}(\mathbb{P}_r, \mathbb{Q}_\theta^{m_i}) = \log 2 > \alpha_0$ , which is in contradiction with our assumption that  $\alpha_0 < \log(2)$ .

We now prove the second assertion namely if  $\alpha = \log 2$  then the minimum common support is zero. Since  $\mathbb{P}_r$  is compactly supported, there exists  $x_1 \in \mathbb{R}^n$ , such that  $\text{dist}(x_1, \text{supp}(\mathbb{P}_r)) > 2$ . Let  $\mathbb{Q}_1$  be a probability distribution on  $B_1(x_1)$ . Then,  $\text{JSD}(\mathbb{P}_r, \mathbb{Q}_1) = \log 2$  and  $\mathcal{L}^k(\text{supp}(\mathbb{P}_r) \cap \text{supp}(\mathbb{Q}_1)) = 0$ . Therefore,  $\text{MCS}_{\log 2} = 0$ .  $\square$

**Theorem 6.1** (Theorem 3.1 in the main paper). *Under the assumptions in Proposition 3.1,  $\text{MCS}^\alpha(\mathbb{P}_r)$  decreases with respect to  $\alpha$  on the interval  $[0, \log 2)$ .*

*Proof.* Let  $\alpha < \beta$  be the two JSD values. By definition, since  $\Omega_\alpha \subset \Omega_\beta$ , we have  $\text{MCS}^\alpha(p_r) \geq \text{MCS}^\beta(p_r)$  automatically.  $\square$

**Theorem 6.2** (Theorem 3.2 in the main paper). *Let  $\epsilon > 0$  and  $\mathbb{P}_r$  be a fixed distributions. Moreover let Assumption A hold. Let  $\Gamma = \{\mathbb{Q}_\theta : W_p(\mathbb{P}_r, \mathbb{Q}_\theta) < \epsilon\}$ . Consider the decomposition:  $\Gamma = \Gamma_1 \cup \Gamma_2$ , where  $\Gamma_1 = \{\mathbb{Q}_\theta : W_p(\mathbb{P}_r, \mathbb{Q}_\theta) < \epsilon; \mu(\text{supp}(\mathbb{P}_r) \cap \text{supp}(\mathbb{Q}_\theta)) > 0\}$  and  $\Gamma_2 = \Gamma - \Gamma_1$ . Then  $\Gamma_2$  is dense in  $\Gamma$ .*

*Proof.* Let  $q_{\theta_0} \in \Gamma_1$  and  $\delta = (\epsilon - W_p(\mathbb{P}_r, \mathbb{Q}_{\theta_0}))/10$ . By general position lemma (Guillemin & Pollack, 2010), for almost every  $t \in \mathbb{R}^n$ ,  $\mathcal{M}_\theta + t$  intersects  $\mathcal{M}_r$  transversally. In particular, for almost every <sup>10</sup>  $t_b \in B_\delta^n(0)$ ,  $\mathcal{M}_{\theta_0} + t_b$  intersects  $\mathcal{M}_r$  transversally. The new probability measure  $\mathbb{Q}_{\theta_0} + t_b$  is identical to  $\mathbb{Q}_{\theta_0}$  except that its support is translated by  $t_b$ . The difference lies in the fact that the common support of the new measure  $\mathbb{Q}_{\theta_0} + t_b$  and  $\mathbb{P}_r$  has measure zero. This translation only affects  $\mathcal{M}_{\theta_0}$  by  $\delta$ , so by definition of  $\delta$

$$W_p(\mathbb{P}_r, \mathbb{Q}_{\theta_0} + t_b) < W_p(\mathbb{P}_r, \mathbb{Q}_{\theta_0}) + \delta < \epsilon$$

by recalling definition of Wasserstein distance. Since we can make  $\delta$  arbitrarily small, we have shown for every  $q_\theta \in \Gamma_1$ , we can find another  $q_{\theta_0} + t_b \in \Gamma_2$  that is as close as we like. This proves the desired claim.  $\square$

<sup>10</sup> Almost every with respect to Lebesgue measure  $\mathcal{L}^n$ .

**Proposition 6.2** (Proposition 4.2 in the main paper). Let  $D^*(x) = \frac{p_r(x)}{q_{\theta_0}(x) + p_r(x)}$  be the optimal discriminator, for  $\theta_0$  fixed. Under **Assumption C** and **Assumption D**, we have:

$$\begin{aligned} & \nabla_\theta \mathbb{E}_{z \sim p(z)} [-\log D^*(g_\theta(z))] |_{\theta=\theta_0} \\ &= 2 \nabla_\theta \text{KL}(\mathbb{Q}_m || \mathbb{P}_r) |_{\theta=\theta_0} \\ &= \mathbb{E}_{\mathbb{Q}_\theta} \left[ \nabla_\theta \log(q_\theta) \left( 1 + \log \left( \frac{q_m}{p_r} \right) \right) \right] \Big|_{\theta=\theta_0}, \quad (7) \end{aligned}$$

and for the standard JSD:

$$\begin{aligned} & \nabla_\theta \mathbb{E}_{z \sim p(z)} [\log(1 - D^*(g_\theta(z)))] |_{\theta=\theta_0} \\ &= \mathbb{E}_{\mathbb{Q}_\theta} \left[ \nabla_\theta \log(q_\theta) \log \left( \frac{q_\theta}{q_m} \right) \right] \Big|_{\theta=\theta_0}. \quad (8) \end{aligned}$$

*Proof.* It is known from [Arjovsky & Bottou \(2017\)](#) that

$$\begin{aligned} & \mathbb{E}_{z \sim p(z)} [-\nabla_\theta \log D^*(g_\theta(z))] |_{\theta=\theta_0} \\ &= \nabla_\theta [\text{KL}(\mathbb{Q}_\theta || \mathbb{P}_r) - 2 \text{JSD}(\mathbb{P}_r, \mathbb{Q}_\theta)] |_{\theta=\theta_0} \end{aligned}$$

By definition of the Kullback Leibler divergence, Jensen Shannon distance and from **Assumption A**

$$\begin{aligned} & \text{KL}(\mathbb{Q}_\theta || \mathbb{P}_r) - 2 \text{JSD}(\mathbb{P}_r, \mathbb{Q}_\theta) \\ &= \text{KL}(\mathbb{Q}_\theta || \mathbb{P}_r) - \text{KL}(\mathbb{P}_r || \mathbb{Q}_m) - \text{KL}(\mathbb{Q}_\theta || \mathbb{Q}_m) \\ &= 2 \text{KL}(\mathbb{Q}_m || \mathbb{P}_r) \end{aligned}$$

Therefore the generator is trained by effectively optimizing the reverse KL between the mixture  $\mathbb{Q}_m$  and the real distribution  $\mathbb{P}_r$ . Hence, using that  $\nabla_\theta q_m(x) = \nabla_\theta q_\theta(x)/2$

$$\begin{aligned} & \nabla_\theta 2 \text{KL}(\mathbb{Q}_m || \mathbb{P}_r) \\ &= \mathbb{E}_{\mathbb{Q}_\theta} \left[ \nabla_\theta \log(q_\theta) + \log \left( \frac{q_m(x)}{p_r(x)} \right) \nabla_\theta \log(q_\theta) \right]. \end{aligned}$$

From [\(Arjovsky & Bottou, 2017\)](#), we know that

$$\begin{aligned} & \mathbb{E}_{z \sim p(z)} [\nabla_\theta \log(1 - D^*(g_\theta(z)))] |_{\theta=\theta_0} \\ &= 2 \nabla_\theta \text{JSD}(\mathbb{Q}_\theta, \mathbb{P}_r) |_{\theta=\theta_0}. \end{aligned}$$

Since

$$\text{KL}(\mathbb{Q}_\theta || \mathbb{P}_r) - 2 \text{JSD}(\mathbb{P}_r, \mathbb{Q}_\theta) = 2 \text{KL}(\mathbb{Q}_m || \mathbb{P}_r),$$

we deduce from Proposition 4.2

$$\begin{aligned} & 2 \nabla_\theta \text{JSD}(\mathbb{P}_r, \mathbb{Q}_\theta) \\ &= \int \nabla_\theta q_\theta(x) \log \left( \frac{q_\theta(x)}{q_m(x)} \right) dx. \end{aligned}$$

□

**Acknowledgement** We would like to thank Joey Bose for his technical support, Gavin Ding for his discussion, and Hamidreza Saghir and Cathal Smyth for their edits and corrections.

## References

- Arjovsky, Martin and Bottou, Léon. Towards principled methods for training generative adversarial networks. In *NIPS 2016 Workshop on Adversarial Training. In review for ICLR*, volume 2016, 2017.
- Arjovsky, Martin, Chintala, Soumith, and Bottou, Léon. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.
- Arora, Sanjeev and Zhang, Yi. Do gans actually learn the distribution? an empirical study. *arXiv preprint arXiv:1706.08224*, 2017.
- Arora, Sanjeev, Ge, Rong, Liang, Yingyu, Ma, Tengyu, and Zhang, Yi. Generalization and equilibrium in generative adversarial nets (gans). *arXiv preprint arXiv:1703.00573*, 2017.
- Bellec, Marc G, Danihelka, Ivo, Dabney, Will, Mohamed, Shakir, Lakshminarayanan, Balaji, Hoyer, Stephan, and Munos, Rémi. The cramer distance as a solution to biased wasserstein gradients. *arXiv preprint arXiv:1705.10743*, 2017.
- Berthelot, David, Schumm, Tom, and Metz, Luke. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017.
- Goodfellow, Ian. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.
- Goodfellow, Ian, Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron, and Bengio, Yoshua. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.
- Guillemin, Victor and Pollack, Alan. *Differential topology*, volume 370. American Mathematical Soc., 2010.
- Gulrajani, Ishaan, Ahmed, Faruk, Arjovsky, Martin, Dumoulin, Vincent, and Courville, Aaron. Improved training of wasserstein gans. *arXiv preprint arXiv:1704.00028*, 2017.
- Kingma, Diederik P and Welling, Max. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- Li, Chun-Liang, Chang, Wei-Cheng, Cheng, Yu, Yang, Yiming, and Póczos, Barnabás. Mmd gan: Towards deeper understanding of moment matching network. *arXiv preprint arXiv:1705.08584*, 2017.
- McCann, Robert J. Polar factorization of maps on riemannian manifolds. *Geometric and Functional Analysis*, 11(3):589–608, 2001.
- McCann, Robert J and Guillen, Nestor. Five lectures on optimal transportation: geometry, regularity and applications. *Analysis and geometry of metric measure spaces: lecture notes of the séminaire de Mathématiques Supérieure (SMS) Montréal*, pp. 145–180, 2011.
- Monge, Gaspard. *Mémoire sur la théorie des déblais et des remblais*. De l’Imprimerie Royale, 1781.
- Narayanan, Hariharan and Mitter, Sanjoy. Sample complexity of testing the manifold hypothesis. In *Advances in Neural Information Processing Systems*, pp. 1786–1794, 2010.
- Nowozin, Sebastian, Cseke, Botond, and Tomioka, Ryota. f-gan: Training generative neural samplers using variational divergence minimization. *arXiv preprint arXiv:1606.00709*, 2016.
- Radford, Alec, Metz, Luke, and Chintala, Soumith. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.
- Villani, Cédric. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.