Why Wasserstein distance is better for training GANs: A summary

Arnab Mondal, School of Computer Science

McGill University, Montreal

Fall, 2019

Term Paper

COMP 599 (Mathematical Techniques for Machine Learning)

Instructor Prof. Prakash Panangaden

Abstract

This term paper explores the effect of different distance metrics in the space of probability distribution on the training of generative models, especially GANs. The first chapter starts with some theoretical insights on the well-known distances used and shows which distance GAN actually minimizes. The second chapter dives into the theoretical understanding of the problems of training GANs and gives some insights into the manifold of the generator probability distribution. The final chapter concludes how Wasserstein distance is theoretically a better metric than other distances. The paper derives heavily from original work on GAN and Wasserstein GAN. Rather than writing the proofs which can be found in the referred articles or providing practical tricks, more focus has been given on establishing the general idea and how we can improve the model mathematically. To maintain the coherence and uniformity of notations, some of the theorems and their proofs might look different than what has been proposed in the original papers.

Table of Contents

Abstract i

1 Generative Adversarial Networks 1

1.1 Distance and divergence between distributions 1

1.2 Losses in Vanilla GAN 4

2 Problems with training GANs 5

2.1 The Perfect discriminator 6

2.2 Vanishing Gradients 9

3 Wasserstein GAN 10

3.1 Optimal Transport and Wasserstein distance 11

3.2 Continuity of Wasserstein distance 13

Acknowledgements 14

Chapter 1

Generative Adversarial Networks

In the last five years, there has been a growing interest in generative modelling of data, and two of the fundamental models in this field are Variational Autoencoders(VAEs) [4] and Generative Adversarial Networks(GANs) [3]. It is interesting to see that the architecture of their generator doesn't change significantly as for both the cases we first sample from a simple prior $z \sim p(z)$ and then output our generated image $g_\theta(z)$ where g_θ is a neural network parameterized by θ. The principal difference lies in how g_θ is trained. Mathematically speaking, we want new samples from real distribution $\mathbb{P}_r$, and the problem we are trying to solve is rather than estimating the density of $\mathbb{P}_r$ which may not exist, we try to make $\mathbb{P}_g$, the distribution of generated samples from g_θ, resemble distribution $\mathbb{P}_r$ as closely as possible by changing θ. For this, we need a notion of closeness between two distributions that we can minimize, which we are going to define in the following section. Note that the ability to easily generate samples is often more useful than knowing the numerical value of the density in most of the practical problems like image super-resolution.

1.1 Distance and divergence between distributions

Before we start defining different distances and divergences, let us define the space we are working on. Let $\mathcal{X}$ be a compact metric set (in this case, the space of images $[0,1]^d$, which has d pixels) and let $\sum$ be the set of all Borel subsets of $\mathcal{X}$ which is essentially the σ-algebra. Now we have $\mathbb{P}_g, \mathbb{P}_r \in \mathcal{M}(\mathcal{X})$, where $\mathcal{M}(\mathcal{X})$ denote the space of probability measures defined on $\mathcal{X}$. Let's also assume $\mathbb{P}_g$ and $\mathbb{P}_r$ to be absolutely continuous with respect to a measure μ defined on $\mathcal{X}$ and therefore have densities $P_r(x)$ and $P_g(x)$. What absolute continuity means is that, $\forall A \in \sum$ if $\mu(A) = 0 \implies \mathbb{P}_r(A) = 0 \quad \& \quad \mathbb{P}_g(A) = 0$ and only then $\forall A \in \sum$ we can write $\mathbb{P}_r(A) = \int_A P_r(x) d\mu(x) \quad \& \quad \mathbb{P}_g(A) = \int_A P_g(x) d\mu(x)$ where $P_r(x)$ and $P_g(x)$ are measurable functions known as probability densities. Now we are ready to define different elementary distances and divergences between $\mathbb{P}_g$ and $\mathbb{P}_r$:

- *Total Variation* (TV) distance :

$$\delta(\mathbb{P}_g, \mathbb{P}_r) = \sup_{A \in \sum} |\mathbb{P}_g(A) - \mathbb{P}_r(A)| \tag{1.1}$$

  which is informally the largest possible difference between the probability that two probability measures can assign to the same event in the σ-algebra.

- *Kullback-Leibler* (KL) divergence :

$$KL(\mathbb{P}_r || \mathbb{P}_g) = \int_{\mathcal{X}} \log(\frac{P_r(x)}{P_g(x)}) P_r(x) d\mu(x) \tag{1.2}$$

  It is interesting to note that it has a unique minimum at $\mathbb{P}_g = \mathbb{P}_r$ and it doesn't require knowledge of unknown $P_r(x)$. The second statement is true because minimizing KL divergence is essentially maximizing the log likelihood of the our data ($x^{(1)}, .., x^{(m)}$) which given by:

$$\max_{\theta} \frac{1}{m} \sum_{m}^{i=1} \log P_\theta(x^{(i)}) \tag{1.3}$$

  where P_θ is the density of a parameterized distribution $\mathbb{P}_\theta$. In our case it is $\mathbb{P}_g$ which depends on the parameters of the generator. This duality can be easily proved by writing:

$$\min \int_{\mathcal{X}} \log(\frac{P_r(x)}{P_g(x)})P_r(x)d\mu(x) \implies \max \int_{\mathcal{X}} \log(P_g(x))P_r(x)d\mu(x)) \tag{1.4}$$

Now all that is left is applying the law of large numbers to the expression we obtained above. It is also interesting to note how this divergence is not symmetrical. In the divergence equation if $P_r(x) > 0$ but $P_g(x) \to 0$, the integrand inside the KL grows to infinity while when $P_r(x) \to 0$ and $P_g(x) > 0$ the KL term goes to 0 meaning the cost function assigns high cost for generator distribution not covering parts of true distribution while it is not efficient when we generate fake-looking sample. Now, if we would minimize $KL(\mathbb{P}_r||\mathbb{P}_g)$, the whole effect would be reversed, and we would pay a high cost for generating fake-looking samples. Note that as VAE focus on maximizing the approximate likelihood of the examples, they share the limitation of the standard model.

- *Jensen-Shannon* (JS) divergence :

  $$JS(\mathbb{P}_r, \mathbb{P}_g) = KL(\mathbb{P}_r||\mathbb{P}_m) + KL(\mathbb{P}_g||\mathbb{P}_m) \tag{1.5}$$

  where $\mathbb{P}_m$ is the average $(\mathbb{P}_r + \mathbb{P}_g)/2$. This provide a middle ground between the two individual cost functions. In the next section we will see how the GANs actually optimize this objective.

Let us wait for the *Wasserstein* distance until the last chapter when we formally define optimal transport.In the next section let's try to get a mathematical outlook of the losses in GAN and their optimal values.

1.2 Losses in Vanilla GAN

This section is derived heavily from the original work on GANs by Goodfellow et al. [3] Before we start let us write the loss of a GAN:

$$\min_{g_\theta} \max_{d_\phi} L(g_\theta, d_\phi) = \mathbb{E}_{x \sim P_r(x)}[\log d_\phi(x)] + \mathbb{E}_{x \sim P_g(x)}[\log(1 - d_\phi(x))] \tag{1.6}$$

here g_θ and d_ϕ are clearly parameterized generator and discriminator. It can be noted that the first term has no impact on the g_θ during optimizer update. Now to obtain the optimal value of the discriminator we need to maximize the following:

$$L(g_\theta, d^*_\phi) = \max_{d_\phi} \int_x (P_r(x) \log d_\phi(x) + P_g(x) \log(1 - d_\phi(x)))dx \tag{1.7}$$

Now it isn't hard to see with a bit of calculus (if we differentiate the term inside the integral and set it to 0) that the best value of the discriminator is $d^*_\phi(x) = \frac{P_r(x)}{P_r(x)+P_g(x)} \in [0, 1]$. Now this leads us to the next question what is the global optimum. We see that when both g_θ and d_ϕ are at their optimal values, we have $P_g(x) = P_r(x)$ and $d^*_\phi(x) = 1/2$ and substituting that in the above equation gives us $L(g^*_\theta, d^*_\phi) = -2 \log 2$. If we write down the expression of $JS(\mathbb{P}_r, \mathbb{P}_g)$ and plug in the value of $L(g_\theta, d^*_\phi) = \int_x (P_r(x) \log \frac{P_r(x)}{P_r(x)+P_g(x)} + P_g(x) \log \frac{P_g(x)}{P_r(x)+P_g(x)})dx$ we get:

$$L(g_\theta, d^*_\phi) = JS(\mathbb{P}_r, \mathbb{P}_g) - 2 \log 2 \tag{1.8}$$

This means that the GAN generator essentially optimizes the JS divergence when the discriminator achieves optimality. The next chapter presents the major problems of training GANs.

Chapter 2

Problems with training GANs

Although GANs have achieved massive success in real looking image generation, training them is not easy. There are multiple reasons which contribute to making it slow and unstable. Below I have listed some of the major problems:

- *Finding Nash equilibrium:* As mention in Salimans et al. [6], when two models are trained simultaneously to find a Nash equilibrium to a two-player non-coperative game, updating the gradients concurrently doesn't guarantee convergence. One beautiful example of this which I found in the paper is: suppose one player minimizes xy with respect to x and another player minimizes $-xy$ with respect to y then gradient descent doesn't converge to $x = y = 0$.

- *Low dimensional support:*Though the dimension of real-world images seem high but their support lies in lower dimensional manifold. This makes the distribution $\mathbb{P}_r$ and $\mathbb{P}_g$ almost certainly disjoint which makes it easy for the discriminator to find its optimum. This leads us to our next problem.

- *Vanishing Gradients:* If the discriminator reaches its optimal solution the gradient of loss function drops down to zero making the learning very slow.

- *Mode collapse:* The generator may get stuck in a setting where it always produces same outputs. Even though it manages to trick the discriminator it fails to represent complex real world image distribution which has high variety.

We are going to deal with the problem of Low dimensional support and vanishing gradients theoretically in this chapter(Note that mode collapse is not being covered in this papers). The next sections would be heavily derived from the excellent theoretical paper by Martin Arjovsky and Leon Bottou [1]. In the next section, we'll see some of the theorems they proposed.

2.1 The Perfect discriminator

From experimental results it has been found that even when the samples are remarkably good and their supports are likely to intersect, the discriminator loss quickly goes to 0 which either imply disjoint supports or the distributions are not absolutely continuous. I want to digress a bit to absolute continuity of random variable to see how it compare with the absolute continuity of measures. The property of a random variable X being absolutely continuous is equivalent to X having a density function $f : \mathcal{X} \to \mathbb{R}$ such that $\mathbb{P}(X \in A) = \int_A f(x)dx$ which is a consequence of Radon-Nikodym theorem. A random variable with support in low dimensional manifold will not be absolutely continuous. This can be seen by taking M a low dimensional manifold to be support of X but since M has 0 Lebesgue measure in $\mathcal{X}$ for it to be absolutely continuous $\mathbb{P}(X \in M) = 0$ which contradicts our initial assumption. There is empirical and theoretical evidence given by Narayanan & Mittter [5] that $\mathbb{P}_r$ is indeed concentrated in low dimensional manifold. Now it can be shown that for well behaved function g_θ we have the support of $\mathbb{P}_g$ in lower dimension give the prior distribution space $\mathcal{Z}$ have lower dimension that $\mathcal{X}$. This can be formally written in the following lemma:

Lemma 2.1. *Let $g_\theta : \mathcal{Z} \to \mathcal{X}$ be a function composed of affine transformations and pointwise nonlinearities which can be either be rectifiers, leaky rectifiers or any smooth strictly increasing* *functions. Then, $g_\theta(\mathcal{Z})$ is contained in a countable union of manifolds of dimension at most dim $\mathcal{Z}$.*

The proof of this lemma is non-trivial, interesting to read and can be found in the appendix A of Arjovsky et al. [1] Now we want to show that given support of $\mathbb{P}_r$ and $\mathbb{P}_g$ are disjoint or lie in low dimensional manifolds, there is always a perfect discriminator between them. Note that our discriminator $d_\phi : \mathcal{X} \rightarrow [0,1]$ is accurate if it takes value 1 on a set that contains the support of $\mathbb{P}_r$ and value 0 on a set that contains the support of $\mathbb{P}_g$. Now we state a very important theorem for the disjoint support case:

Theorem 2.1. *If two distributions $\mathbb{P}_r$ and $\mathbb{P}_g$ have support contained on two disjoint compact subsets M_r and M_g respectively, then there is a smooth optimal discrimator $d_\phi^* : \mathcal{X} \rightarrow [0,1]$ that has accuracy 1 and $\nabla_x d_\phi^*(x) = 0 \; \forall x \in M_r \cup M_g$.*

Proof. Let us consider $d(M_r, M_g) = \epsilon > 0$ be the distance between two sets as M_r and M_g are disjoint & compact. We can construct two other compact and disjoint sets $M_r^{'} = \{x : d(x, M_r) \leq \epsilon/3\}$ and $M_g^{'} = \{x : d(x, M_g) \leq \epsilon/3\}$. Note that we could have chosen anything $> \epsilon/2$ for the margin. By Urysohn's smooth lemma there exist a smooth function $d_\phi^* : \mathcal{X} \rightarrow [0,1]$ such that $d_\phi^*(x)|_{x \in M_r^{'}} = 1$ and $d_\phi^*(x)|_{x \in M_g^{'}} = 0$. For the second part let us consider an open ball $B_{\epsilon/3}(x)$ around $x \in M_r$ which gives $d_\phi^*(x)|_{x \in B_{\epsilon/3}(x)} = 1$ and hence $\nabla_x d_\phi^*(x) = 0$. Analogously for $x \in M_g$. □

Our goal now is to show that when two manifold have lower dimensional support and they intersect we still have a perfect discriminator. But before that, we look into some definitions and Lemmas of the alignment of manifolds. I have directly taken them from Arjovsky et al. [1] and formulated in my way for our setting. The proofs of Lemma 2.2 and 2.3 can be found in the Appendix A of their paper. An important thing to note which the authors haven't clarified properly is that the manifolds we are dealing with here are assumed to be smooth differentiable manifolds which may not be the case for supports of real world data. Nevertheless it gives us convenience to prove things.

Definition 2.1. Let M_1 and M_2 be two boundary free regular submanifolds of $\mathcal{X}$. Let $x \in$

$M_1 \cap M_2$ be an intersection point of the two manifolds. M_1 and M_2 intersect transversally in x if $\mathcal{T}_x M_1 + \mathcal{T}_x M_2 = \mathcal{T}_x \mathcal{X}$, where $\mathcal{T}_x M$ means the tangent space of M around x.

Definition 2.2. Two manifolds without boundary M_1 and M_2 perfectly align if there is an $x \in M_1 \cap M_2$ such that M_1 and M_2 don't intersect transversally in x. Note that two manifolds M_1 and M_2 (with or without boundary) perfectly align if any of the boundary free manifold pairs (Int M_1, Int M_2), (Int M_1, ∂M_2), (∂M_1, Int M_2) or (∂M_1, ∂M_2) perfectly align where the boundary and interior of a manifold M by ∂M and Int M respectively.

***Lemma* 2.2.** *Let M_1 and M_2 be two regular submanifolds of $\mathcal{X}$ that don't have full dimension. Let η_1 and η_2 be arbitrary independent continuous random variables. Therefore define the perturbed manifolds as $\tilde{M_1} = M_1 + \eta_1$ and $\tilde{M_2} = M_2 + \eta_2$. Then*

$$\mathbb{P}_{\eta_1,\eta_2}(\tilde{M_1}\, doesn't\, perfectly\, align\, with\, \tilde{M_2}) = 1$$

***Lemma* 2.3.** *Let M_1 and M_2 be two regular submanifolds of $\mathcal{X}$ that don't have full dimension and don't perfectly align. . Let $\mathcal{L} = M_1 \cap M_2$. If M_1 and M_2 don't have boundary, then $\mathcal{L}$ is also a manifold, and has strictly lower dimension than both M_1 and M_2. If they have boundary, $\mathcal{L}$ is a union of at most 4 strictly lower dimensional manifolds. In both cases, $\mathcal{L}$ has measure 0 in both M_1 and M_2.*

Informally Lemma 2.2 and 2.3 tells us that we can safely assume that in practice two lower dimensional submanifold of a higher dimensional space never perfectly align which results in their intersection be in further lower dimension and have a measure 0. We are going to use it in the last theorem of this section about existence of perfect discriminator.

Theorem 2.2. *Let $\mathbb{P}_r$ and $\mathbb{P}_g$ be two distributions that have support contained in two closed manifolds M_r and M_g that don't perfectly align and don't have full dimension. We further assume that $\mathbb{P}_r$ and $\mathbb{P}_g$ are absolutely continuous in their respective manifolds. Then, there exists an optimal discriminator $d^*_\phi : \mathcal{X} \to [0, 1]$ that has accuracy 1 and for almost any x in M_r or M_g, d^*_ϕ is smooth in a neighbourhood of x and $\nabla_x d^*_\phi(x) = 0$*

Proof. Clearly by absolute continuity we have $\mathbb{P}_r(\mathcal{L}) = 0$ and $\mathbb{P}_g(\mathcal{L}) = 0$ which implies support of $\mathbb{P}_r$ is in $M_r \setminus \mathcal{L}$ and that of $\mathbb{P}_g$ is in $M_g \setminus \mathcal{L}$. Now as $\mathcal{X} \setminus M_g$ is an open set, for every $x \in M_r \setminus \mathcal{L}$ there exist a ball of radius ϵ_x such that $B_{\epsilon_x}(x) \cap M_g = \phi$. Let us define $\hat{M}_r = \bigcup_{x \in M_r \setminus \mathcal{L}} B_{\epsilon_x/3}(x)$ and $\hat{M}_g$ analogously. By construction they are open sets in $\mathcal{X}$ and support of $\mathbb{P}_r$ & $\mathbb{P}_g$ are in them. Also notice that $\hat{M}_r \cap \hat{M}_p = \phi$. Now we can use similar reasoning as in theorem 2.1 and prove the rest. □

2.2 Vanishing Gradients

In the previous section we saw that how the optimal discriminator becomes perfect which makes its gradient 0 almost everywhere. In this section we want to see what happens when we pass the gradients from discriminator to the generator. The following theorem makes it clear and note that $||d_\phi|| = \sup_{x \in \mathcal{X}} |d_\phi(x)| + ||\nabla_x d_\phi(x)||_2$:

Theorem 2.3. *If the conditions of Theorem 2.1 or Theorem 2.2 are satisfied,* $||d_\phi - d^*_\phi|| < \epsilon$ *and* $\mathbb{E}_{z \sim p(z)}[||J_\theta g_\theta(z)||_2^2] \leq k$ *where* J_θ *is the jacobian, then:*

$$||\nabla_\theta \mathbb{E}_{z \sim p(z)}[\log(1 - d_\phi(g_\theta(z)))]||_2 < k\frac{\epsilon}{1-\epsilon} \tag{2.1}$$

The proof is straightforward using Jensen's inequality & chain rule which can be found in [1]. Note the condition $\mathbb{E}_{z \sim p(z)}[||J_\theta g_\theta(z)||_2^2] \leq k$ is trivially verified for a uniform prior and neural network based generator where k would depend on θ. To show that the gradients of the generator vanishes as discriminator approaches it's optimum we take $||d_\phi - d^*_\phi|| \rightarrow 0$:

$$\lim_{||d_\phi - d^*_\phi|| \rightarrow 0} \nabla_\theta \mathbb{E}_{z \sim p(z)}[\log(1 - d_\phi(g_\theta(z)))] = 0 \tag{2.2}$$

Researcher have also tried taking gradient of $\nabla_\theta \mathbb{E}_{z \sim p(z)}[-\log(d_\phi(g_\theta(z)))]$ but it's mathematically not a good option either as shown in Theorem 2.5 and 2.6 in [1].

Chapter 3

Wasserstein GAN

In the last chapter, we have seen all the theoretical problems of GAN training resulting from a lower-dimensional manifold of distributions and their disjoint support. One way to tackle this problem is to artificially spread out the distribution and create higher chances for two probability distributions to have overlap by adding continuous noises onto the inputs of the discriminator d_ϕ. This works well in practice. Though there are many practical techniques that can be incorporated for training GANs as discussed in Saliman et al. [6] but we are more interested in dealing with the distance between distributions side of thing, as discussed a bit in chapter 1. As beautifully shown through a toy example of learning parallel lines in Arjovsky et al. [2] that when the supports are disjoint or are in lower-dimensional manifold the TV distance, KL and JS divergences are maxed out, discontinuous and primarily the target distribution cannot be learned from them using gradients descent. This is what tells us that we need a softer metric and also some notion of distance between points in manifolds. Before I introduce the Wasserstein metric, let us take a brief look at the crucial concepts of transport theory.

3.1 Optimal Transport and Wasserstein distance

Below I am going to present the Monge's Problem and how Kantorovich changed it. Let us define the setting for our problem. Let $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ be two probability spaces and $c : [X, Y] \to [0, +\infty]$ be a cost function. Note that the meaning of the space denoted by $\mathcal{X}$ has changed in this section.

Monge's Problem *The problem asks us to find a measurable Transport function $T : \mathcal{X} \to \mathcal{Y}$ such that $\nu = \mu \circ T^{-1}$ is the pushforward measure and $\int_{\mathcal{X}} c(x, T(x)) d\mu$ is minimized.* This is equivalent to transport a mass from one distribution to another with a cost assigned for each unit of mass transported. Kantorovich came up with an always solvable version of this problem where he replaced the transport function with transport plan. Next we see the definition of coupling and Monge-Kantorovich's Problem.

Coupling*. Coupling μ and ν means constructing two random variables X and Y on some probability space $(\sigma, \mathbb{P})$ in such a way that law(X) = μ and law(X) = ν. Both the couple (X, Y) and law of (X, Y) are called coupling of (μ, ν). Without loss of generality one can choose $\sigma = \mathcal{X} \times \mathcal{Y}$ such that coupling μ and ν means constructing a measure π on $\mathcal{X} \times \mathcal{Y}$ such that π have μ and ν as marginals on $\mathcal{X}$ and $\mathcal{Y}$.*

Monge-Kantorovich's Problem *The problem asks us to find a transport plan instead, which is essentially a coupling $\pi \in \Pi(\mu, \nu)$, where Π denote the set of all coupling, such that $\mathcal{K}(\pi) = \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\pi$ is minimized.*

In Polish probability spaces(i.e. complete, separable, metric) $\mathcal{X}$ of images $[0, 1]^d$ with a metric $\mathcal{D}$ and two probability measures $\mathbb{P}_r$ and $\mathbb{P}_g$, the Earth-Mover(EM) or Wasserstein-1 distance is defined as:

$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\pi \in \Pi(\mathbb{P}_r, \mathbb{P}_g)} \int_{\mathcal{X} \times \mathcal{X}} \mathcal{D}(x, y) d\pi \tag{3.1}$$

where $\Pi(\mathbb{P}_r, \mathbb{P}_g)$ denotes the set of all joint distributions π whose marginals are $\mathbb{P}_r$ and $\mathbb{P}_g$. This distance is basically the cost of optimal transport plan. Also in such well behaved spaces, the Kantorovich Duality(which is essentially a dual problem we solve and can be found in details in [7]) reduces to Kantorovich-Rubenstein duality (This is the link to an amazing post where the author uses numerical methods to formulate and show the duality in discrete probability space and proved it in continuous space. It's worth a read.):

$$\inf_{\pi \in \Pi(\mathbb{P}_r, \mathbb{P}_g)} \int_{\mathcal{X} \times \mathcal{X}} \mathcal{D}(x, y) d\pi = \sup_{||f||_L \leq 1} (\int_{\mathcal{X}} f d\mathbb{P}_r - \int_{\mathcal{X}} f d\mathbb{P}_g) \tag{3.2}$$

As infimum in the left hand side of above equation is highly intractable so obtaining the sup on the right hand side is a better solution. Now note that the function $f : \mathcal{X} \rightarrow \mathbb{R}$ is 1-Lipschitz. A real valued function is called K-Lipschitz continuous if there exist a real constant $K \geq 0$ such that $\forall x_1, x_2 \in \mathbb{R}, |f(x_1) - f(x_2)| \leq K|x_1 - x_2|$. Now we restrict the function to K-Lipschitz continuous then we end maximizing $K \times W(\mathbb{P}_r, \mathbb{P}_g)$. Now let, $f_{w_{w \in \mathcal{W}}}$ be the family of all K-Lipschitz for some K, then our problem boils down to:

$$\max_{w \in \mathcal{W}} \mathbb{E}_{x \sim \mathbb{P}_r}[f_w(x)] - \mathbb{E}_{z \sim p(z)}[f_w(g_\theta(z))] \tag{3.3}$$

If the above expression achieves maximum for some $w \in \mathcal{W}$, it would yield a calculation of $W(\mathbb{P}_r, \mathbb{P}_g)$ upto a multiplicative constant. Now if we differentiate $W(\mathbb{P}_r, \mathbb{P}_g)$ with respect to θ and backpropagate through the generator to minimize the distance that should hopefully lead us to the generator optimum. It has been shown by Theorem 3 of Arjovsky et al. [2] that indeed there exists a solution of f and the gradient of the loss with respect to generator parameters is well defined there. Note that to maintain K-Lipschitz continuity of f_w during the training Arjovsky proposed an impressive practical trick that is after gradient update clamp the weight w to a small window resulting in compact parameter space $\mathcal{W}$ and thus f_w gets bounded. Still it suffers unstable training because of the weight clipping which result in slow convergence or vanishing gradient depending on the size of the window. The next section I'll discuss and conclude why the Wasserstein distance is better than other distances.

3.2 Continuity of Wasserstein distance

Let $\mathbb{P}_{g_\theta}$ be the distribution we are trying to learn by optimizing the parameter θ of g_θ. Our entire problem of making generative models converge with tradition KL divergence or JSD lies with the fact that, given the low dimensional support of $\mathbb{P}_r$, the distance or divergence is maxed out almost everywhere in $\mathcal{X}$ and jumps to minima 0 when $\mathbb{P}_{g_\theta} = \mathbb{P}_r$. But what we want is continuity in the mapping $\theta \to \mathbb{P}_{g_\theta}$ because it is equivalent to making the loss $\rho(\mathbb{P}_{g_\theta}, \mathbb{P}_r)$ continuous where ρ is the metric in the space of distributions. Then we can ideally run gradient descent and converge to the optimum. Note that this distance ρ should preferably induce a weaker topolgy. In the appendix A and proof of Theorem 2 of Arjovsky et al. [2] there is an outstanding discussion on how KL induce the strongest topology, followed by JSD and TV, and Wasserstein distance induces the weakest. Now I am going to state the theorem of continuity and differentiability of Wasserstein distance and an important assumption that we need for differentiability. The proofs are given in [2]:

Assumption 3.1. *Let $g : \mathcal{Z} \times \mathbb{R}^{d_\theta} \to \mathcal{X}$ be locally Lipschitz. We say g satisfies the assumption for a certain probability distribution p over $\mathcal{Z}$ if there are local Lipschitz constants $L(\theta, z)$ such that $\mathbb{E}_{z \in p}[L(\theta, z)] < +\infty$.*

Theorem 3.1. *Let $g : \mathcal{Z} \times \mathbb{R}^{d_\theta} \to \mathcal{X}$ be a function, that will be denoted by $g_\theta(z)$ with z the first coordinate and θ the second.*

1. If g is continuous in θ, so is $W(\mathbb{P}_r, \mathbb{P}_{g_\theta})$.

2. If g is locally Lipschitz and satisfies the assumption 3.1, then $W(\mathbb{P}_r, \mathbb{P}_{g_\theta})$ is continuous everywhere, and differentiable almost everywhere.

3. Statements 1-2 are false for JS and all KLs divergences.

Now Arjovsky showed in the paper that for any feed forward neural network g_θ follow assumption 3.1 and hence the $W(\mathbb{P}_r, \mathbb{P}_{g_\theta})$ for them is continuous everywhere, and differentiable almost everywhere. This completes our understanding of why theoretically using Wasserstein distance is better for generative learning of distribution.

Acknowledgements

I want to thank Prof. Prakash Panangaden for teaching COMP 599, which covers all the necessary background needed to understand the relevant papers and to write this term paper. I also want to thank him for his valuable discussion on this project idea. I also want to thank Florence for all the help with assignments, which built my understanding of the subject.

Bibliography

[1] Arjovsky, M., and Bottou, L. Towards principled methods for training generative adversarial networks. arxiv.

[2] Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein gan. *arXiv preprint arXiv:1701.07875* (2017).

[3] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems* (2014), pp. 2672–2680.

[4] Kingma, D. P., and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[5] Narayanan, H., and Mitter, S. Sample complexity of testing the manifold hypothesis. In *Advances in Neural Information Processing Systems* (2010), pp. 1786–1794.

[6] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training gans. In *Advances in neural information processing systems* (2016), pp. 2234–2242.

[7] Villani, C. *Optimal transport: old and new*, vol. 338. Springer Science & Business Media, 2008.