Wasserstein Generative Adversarial Networks

Martin Arjovsky [1] **Soumith Chintala** [2] **Léon Bottou** [1 2]

Abstract

We introduce a new algorithm named WGAN, an alternative to traditional GAN training. In this new model, we show that we can improve the stability of learning, get rid of problems like mode collapse, and provide meaningful learning curves useful for debugging and hyperparameter searches. Furthermore, we show that the corresponding optimization problem is sound, and provide extensive theoretical work highlighting the deep connections to different distances between distributions.

1. Introduction

The problem this paper is concerned with is that of unsupervised learning. Mainly, what does it mean to learn a probability distribution? The classical answer to this is to learn a probability density. This is often done by defining a parametric family of densities $(P_\theta)_{\theta \in \mathbb{R}^d}$ and finding the one that maximized the likelihood on our data: if we have real data examples $\{x^{(i)}\}_{i=1}^m$, we would solve the problem

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

If the real data distribution $\mathbb{P}_r$ admits a density and $\mathbb{P}_\theta$ is the distribution of the parametrized density P_θ, then, asymptotically, this amounts to minimizing the Kullback-Leibler divergence $KL(\mathbb{P}_r \| \mathbb{P}_\theta)$.

For this to make sense, we need the model density P_θ to exist. This is not the case in the rather common situation where we are dealing with distributions supported by low dimensional manifolds. It is then unlikely that the model manifold and the true distribution's support have a non-negligible intersection (see (Arjovsky & Bottou, 2017)), and this means that the KL distance is not defined (or simply infinite).

The typical remedy is to add a noise term to the model distribution. This is why virtually all generative models described in the classical machine learning literature include a noise component. In the simplest case, one assumes a Gaussian noise with relatively high bandwidth in order to cover all the examples. It is well known, for instance, that in the case of image generation models, this noise degrades the quality of the samples and makes them blurry. For example, we can see in the recent paper (Wu et al., 2016) that the optimal standard deviation of the noise added to the model when maximizing likelihood is around 0.1 to each pixel in a generated image, when the pixels were already normalized to be in the range $[0, 1]$. This is a very high amount of noise, so much that when papers report the samples of their models, they don't add the noise term on which they report likelihood numbers. In other words, the added noise term is clearly incorrect for the problem, but is needed to make the maximum likelihood approach work.

Rather than estimating the density of $\mathbb{P}_r$ which may not exist, we can define a random variable Z with a fixed distribution $p(z)$ and pass it through a parametric function $g_\theta : \mathcal{Z} \to \mathcal{X}$ (typically a neural network of some kind) that directly generates samples following a certain distribution $\mathbb{P}_\theta$. By varying θ, we can change this distribution and make it close to the real data distribution $\mathbb{P}_r$. This is useful in two ways. First of all, unlike densities, this approach can represent distributions confined to a low dimensional manifold. Second, the ability to easily generate samples is often more useful than knowing the numerical value of the density (for example in image superresolution or semantic segmentation when considering the conditional distribution of the output image given the input image). In general, it is computationally difficult to generate samples given an arbitrary high dimensional density (Neal, 2001).

Variational Auto-Encoders (VAEs) (Kingma & Welling, 2013) and Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) are well known examples of this approach. Because VAEs focus on the approximate likelihood of the examples, they share the limitation of the standard models and need to fiddle with additional noise terms. GANs offer much more flexibility in the definition of the objective function, including Jensen-Shannon (Goodfellow et al., 2014), and all f-divergences (Nowozin et al., 2016) as well as some exotic combinations (Huszar, 2015). On

[1]Courant Institute of Mathematical Sciences, NY [2]Facebook AI Research, NY. Correspondence to: Martin Arjovsky <martinarjovsky@gmail.com>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, 2017. JMLR: W&CP.

the other hand, training GANs is well known for being delicate and unstable, for reasons theoretically investigated in (Arjovsky & Bottou, 2017).

In this paper, we direct our attention on the various ways to measure how close the model distribution and the real distribution are, or equivalently, on the various ways to define a distance or divergence $\rho(\mathbb{P}_\theta, \mathbb{P}_r)$. The most fundamental difference between such distances is their impact on the convergence of sequences of probability distributions. A sequence of distributions $(\mathbb{P}_t)_{t\in\mathbb{N}}$ converges if and only if there is a distribution $\mathbb{P}_\infty$ such that $\rho(\mathbb{P}_t, \mathbb{P}_\infty)$ tends to zero, something that depends on how exactly the distance ρ is defined. Informally, a distance ρ induces a weaker topology when it makes it easier for a sequence of distribution to converge.[1] Section 2 clarifies how popular probability distances differ in that respect.

In order to optimize the parameter θ, it is of course desirable to define our model distribution $\mathbb{P}_\theta$ in a manner that makes the mapping $\theta \mapsto \mathbb{P}_\theta$ continuous. Continuity means that when a sequence of parameters θ_t converges to θ, the distributions $\mathbb{P}_{\theta_t}$ also converge to $\mathbb{P}_\theta$. However, it is essential to remember that the notion of the convergence of the distributions $\mathbb{P}_{\theta_t}$ depends on the way we compute the distance between distributions. The weaker this distance, the easier it is to define a continuous mapping from θ-space to $\mathbb{P}_\theta$-space, since it's easier for the distributions to converge. The main reason we care about the mapping $\theta \mapsto \mathbb{P}_\theta$ to be continuous is as follows. If ρ is our notion of distance between two distributions, we would like to have a loss function $\theta \mapsto \rho(\mathbb{P}_\theta, \mathbb{P}_r)$ that is continuous, and this is equivalent to having the mapping $\theta \mapsto \mathbb{P}_\theta$ be continuous when using the distance between distributions ρ.

The contributions of this paper are:

- In Section 2, we provide a comprehensive theoretical analysis of how the Earth Mover (EM) distance behaves in comparison to popular probability distances and divergences used in the context of learning distributions.
- In Section 3, we define a form of GAN called Wasserstein-GAN that minimizes a reasonable and efficient approximation of the EM distance, and we theoretically show that the corresponding optimization problem is sound.
- In Section 4, we empirically show that WGANs cure the main training problems of GANs. In particular, training WGANs does not require maintaining a careful balance in training of the discriminator and the generator, does not require a careful design of the network architecture either, and also reduces the mode dropping that is typical in GANs. One of the most compelling practical benefits of WGANs is the ability to continuously estimate the EM distance by training the discriminator to optimality. Because they correlate well with the observed sample quality, plotting these learning curves is very useful for debugging and hyperparameter searches.

2. Different Distances

We now introduce our notation. Let $\mathcal{X}$ be a compact metric set, say the space of images $[0,1]^d$, and let Σ denote the set of all the Borel subsets of $\mathcal{X}$. Let $\text{Prob}(\mathcal{X})$ denote the space of probability measures defined on $\mathcal{X}$. We can now define elementary distances and divergences between two distributions $\mathbb{P}_r, \mathbb{P}_g \in \text{Prob}(\mathcal{X})$:

- The *Total Variation* (TV) distance
$$\delta(\mathbb{P}_r, \mathbb{P}_g) = \sup_{A\in\Sigma} |\mathbb{P}_r(A) - \mathbb{P}_g(A)| \; .$$
- The *Kullback-Leibler* (KL) divergence
$$KL(\mathbb{P}_r \| \mathbb{P}_g) = \int \log\left(\frac{P_r(x)}{P_g(x)}\right) P_r(x) d\mu(x) \; ,$$
where both $\mathbb{P}_r$ and $\mathbb{P}_g$ are assumed to admit densities with respect to a same measure μ defined on $\mathcal{X}$.[2] The KL divergence is famously assymetric and possibly infinite when there are points such that $P_g(x) = 0$ and $P_r(x) > 0$.
- The *Jensen-Shannon* (JS) divergence
$$JS(\mathbb{P}_r, \mathbb{P}_g) = KL(\mathbb{P}_r \| \mathbb{P}_m) + KL(\mathbb{P}_g \| \mathbb{P}_m) \; ,$$
where $\mathbb{P}_m$ is the mixture $(\mathbb{P}_r + \mathbb{P}_g)/2$. This divergence is symmetrical and always defined because we can choose $\mu = \mathbb{P}_m$.
- The *Earth-Mover* (EM) distance or Wasserstein-1
$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma\in\Pi(\mathbb{P}_r,\mathbb{P}_g)} \mathbb{E}_{(x,y)\sim\gamma}\big[\, \|x-y\| \,\big] \; , \quad (1)$$
where $\Pi(\mathbb{P}_r, \mathbb{P}_g)$ is the set of all joint distributions $\gamma(x,y)$ whose marginals are respectively $\mathbb{P}_r$ and $\mathbb{P}_g$. Intuitively, $\gamma(x,y)$ indicates how much "mass" must be transported from x to y in order to transform the distributions $\mathbb{P}_r$ into the distribution $\mathbb{P}_g$. The EM distance then is the "cost" of the optimal transport plan.

[1] More exactly, the topology induced by ρ is weaker than that induced by ρ' when the set of convergent sequences under ρ is a superset of that under ρ'.

[2] Recall that a probability distribution $\mathbb{P}_r \in \text{Prob}(\mathcal{X})$ admits a density $P_r(x)$ with respect to μ, that is, $\forall A \in \Sigma$, $\mathbb{P}_r(A) = \int_A P_r(x) d\mu(x)$, if and only it is absolutely continuous with respect to μ, that is, $\forall A \in \Sigma$, $\mu(A) = 0 \Rightarrow \mathbb{P}_r(A) = 0$.

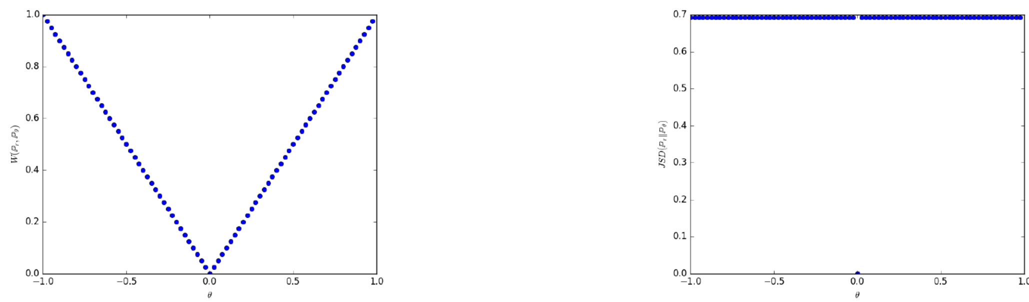

Figure 1: These plots show $\rho(\mathbb{P}_\theta, \mathbb{P}_0)$ as a function of θ when ρ is the EM distance (left plot) or the JS divergence (right plot). The EM plot is continuous and provides a usable gradient everywhere. The JS plot is not continuous and does not provide a usable gradient.

The following example illustrates how apparently simple sequences of probability distributions converge under the EM distance but do not converge under the other distances and divergences defined above.

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

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

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

Example 1 gives a case where we can learn a probability distribution over a low dimensional manifold by doing gradient descent on the EM distance. This cannot be done with the other distances and divergences because the resulting loss function is not even continuous. Although this simple example features distributions with disjoint supports, the same conclusion holds when the supports have a non empty intersection contained in a set of measure zero. This happens to be the case when two low dimensional manifolds intersect in general position (Arjovsky & Bottou, 2017).

Since the Wasserstein distance is much weaker than the JS distance,[3] we can now ask whether $W(\mathbb{P}_r, \mathbb{P}_\theta)$ is a continuous loss function on θ under mild assumptions:

Theorem 1. *Let $\mathbb{P}_r$ be a fixed distribution over $\mathcal{X}$. Let Z be a random variable (e.g Gaussian) over another space $\mathcal{Z}$. Let $\mathbb{P}_\theta$ denote the distribution of $g_\theta(Z)$ where $g : (z, \theta) \in \mathcal{Z} \times \mathbb{R}^d \mapsto g_\theta(z) \in \mathcal{X}$. Then,*

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

As a consequence, learning by minimizing the EM distance makes sense (at least in theory) for neural networks:

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

Both proofs are given in Appendix C.

All this indicates that EM is a much more sensible cost function for our problem than at least the Jensen-Shannon divergence. The following theorem describes the relative strength of the topologies induced by these distances and divergences, with KL the strongest, followed by JS and TV, and EM the weakest.

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

[3] Appendix A explains to the mathematically inclined reader *why* this happens and how we arrived to the idea that Wasserstein is what we should really be optimizing.

[4] By a feedforward neural network we mean a function composed of affine transformations and componentwise Lipschitz nonlinearities (such as the sigmoid, tanh, elu, softplus, etc). A similar but more technical proof is required for ReLUs.

1. *The following statements are equivalent*
   - $\delta(\mathbb{P}_n, \mathbb{P}) \to 0$ *with* δ *the total variation distance.*
   - $JS(\mathbb{P}_n, \mathbb{P}) \to 0$ *with* JS *the Jensen-Shannon divergence.*
2. *The following statements are equivalent*
   - $W(\mathbb{P}_n, \mathbb{P}) \to 0$.
   - $\mathbb{P}_n \xrightarrow{\mathcal{D}} \mathbb{P}$ *where* $\xrightarrow{\mathcal{D}}$ *represents convergence in distribution for random variables.*
3. $KL(\mathbb{P}_n \| \mathbb{P}) \to 0$ *or* $KL(\mathbb{P} \| \mathbb{P}_n) \to 0$ *imply the statements in (1).*
4. *The statements in (1) imply the statements in (2).*

Proof. See Appendix C □

This highlights the fact that the KL, JS, and TV distances are not sensible cost functions when learning distributions supported by low dimensional manifolds. However the EM distance is sensible in that setup. This leads us to the next section where we introduce a practical approximation of optimizing the EM distance.

3. Wasserstein GAN

Again, Theorem 2 points to the fact that $W(\mathbb{P}_r, \mathbb{P}_\theta)$ might have nicer properties when optimized than $JS(\mathbb{P}_r, \mathbb{P}_\theta)$. However, the infimum in (1) is highly intractable. On the other hand, the Kantorovich-Rubinstein duality (Villani, 2009) tells us that

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

where the supremum is over all the 1-Lipschitz functions $f : \mathcal{X} \to \mathbb{R}$. Note that if we replace $\|f\|_L \leq 1$ for $\|f\|_L \leq K$ (consider K-Lipschitz for some constant K), then we end up with $K \cdot W(\mathbb{P}_r, \mathbb{P}_g)$. Therefore, if we have a parameterized family of functions $\{f_w\}_{w\in\mathcal{W}}$ that are all K-Lipschitz for some K, we could consider solving the problem

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

and if the supremum in (2) is attained for some $w \in \mathcal{W}$ (a pretty strong assumption akin to what's assumed when proving consistency of an estimator), this process would yield a calculation of $W(\mathbb{P}_r, \mathbb{P}_\theta)$ up to a multiplicative constant. Furthermore, we could consider differentiating $W(\mathbb{P}_r, \mathbb{P}_\theta)$ (again, up to a constant) by back-proping through equation (2) via estimating $\mathbb{E}_{z\sim p(z)}[\nabla_\theta f_w(g_\theta(z))]$. While this is all intuition, we now prove that this process is principled under the optimality assumption.

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 $t = 0, ..., n_{\text{critic}}$ **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

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

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

and we have

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

when both terms are well-defined.

Proof. See Appendix C □

Now comes the question of finding the function f that solves the maximization problem in equation (2). To roughly approximate this, something that we can do is train a neural network parameterized with weights w lying in a compact space $\mathcal{W}$ and then backprop through $\mathbb{E}_{z\sim p(z)}[\nabla_\theta f_w(g_\theta(z))]$, as we would do with a typical GAN. Note that the fact that $\mathcal{W}$ is compact implies that all the functions f_w will be K-Lipschitz for some K that only depends on $\mathcal{W}$ and not the individual weights, therefore approximating (2) up to an irrelevant scaling factor and the capacity of the 'critic' f_w. In order to have parameters w lie in a compact space, something simple we can do is clamp the weights to a fixed box (say $\mathcal{W} = [-0.01, 0.01]^l$) after each gradient update. The Wasserstein Generative Adversarial Network (WGAN) procedure is described in Algorithm 1.

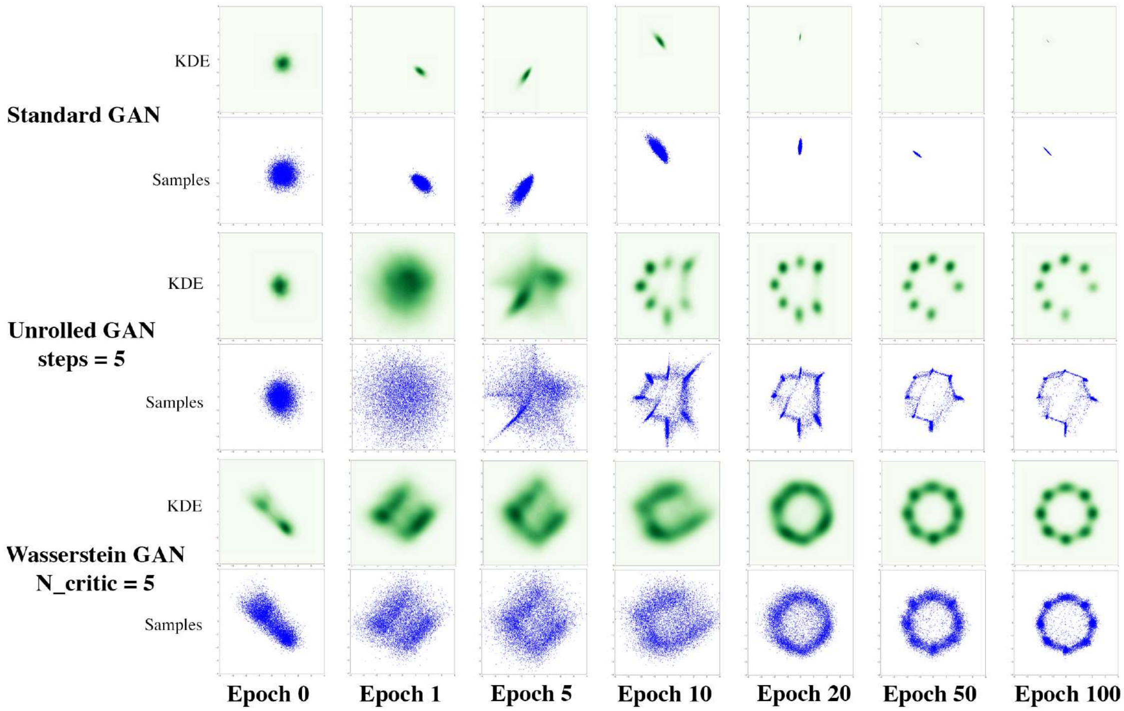

Figure 2: Different methods learning a mixture of 8 gaussians spread in a circle. WGAN is able to learn the distribution without mode collapse. An interesting fact is that the WGAN (much like the Wasserstein distance) seems to capture first the low dimensional structure of the data (the approximate circle) before matching the specific bumps in the density. Green: KDE plots. Blue: samples from the model.

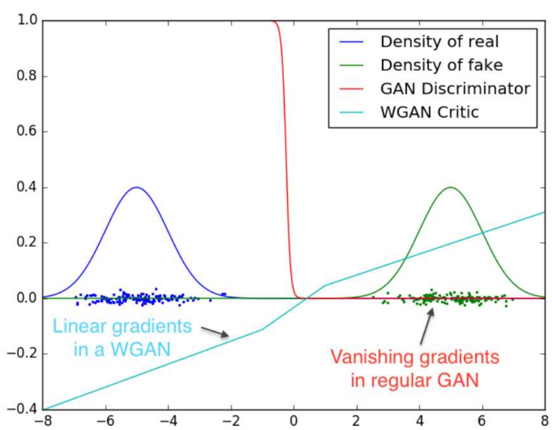

Figure 3: Optimal discriminator and critic when learning to differentiate two Gaussians. As we can see, the traditional GAN discriminator saturates and results in vanishing gradients. Our WGAN critic provides very clean gradients on all parts of the space.

The fact that the EM distance is continuous and differentiable a.e. means that we can (and should) train the critic till optimality. The argument is simple, the more we train the critic, the more reliable gradient of the Wasserstein we get, which is actually useful by the fact that Wasserstein is differentiable almost everywhere. For the JS, as the discriminator gets better the gradients get more reliable but the true gradient is 0 since the JS is locally saturated and we get vanishing gradients, as can be seen in Figure 1 of this paper and Theorem 2.4 of (Arjovsky & Bottou, 2017). In Figure 3 we show a proof of concept of this, where we train a GAN discriminator and a WGAN critic till optimality. The discriminator learns very quickly to distinguish between fake and real, and as expected provides no reliable gradient information. The critic, however, can't saturate, and converges to a linear function that gives remarkably clean gradients everywhere. The fact that we constrain the weights limits the possible growth of the function to be at most linear in different parts of the space, forcing the optimal critic to have this behaviour.

Perhaps more importantly, the fact that we can train the critic till optimality makes it impossible to collapse modes when we do. This is due to the fact that mode collapse comes from the fact that the optimal generator for a *fixed* discriminator is a sum of deltas on the points the discriminator assigns the highest values, as observed by (Goodfellow et al., 2014) and highlighted in (Metz et al., 2016).

In the following section we display the practical benefits of our new algorithm, and we provide an in-depth comparison of its behaviour and that of traditional GANs.

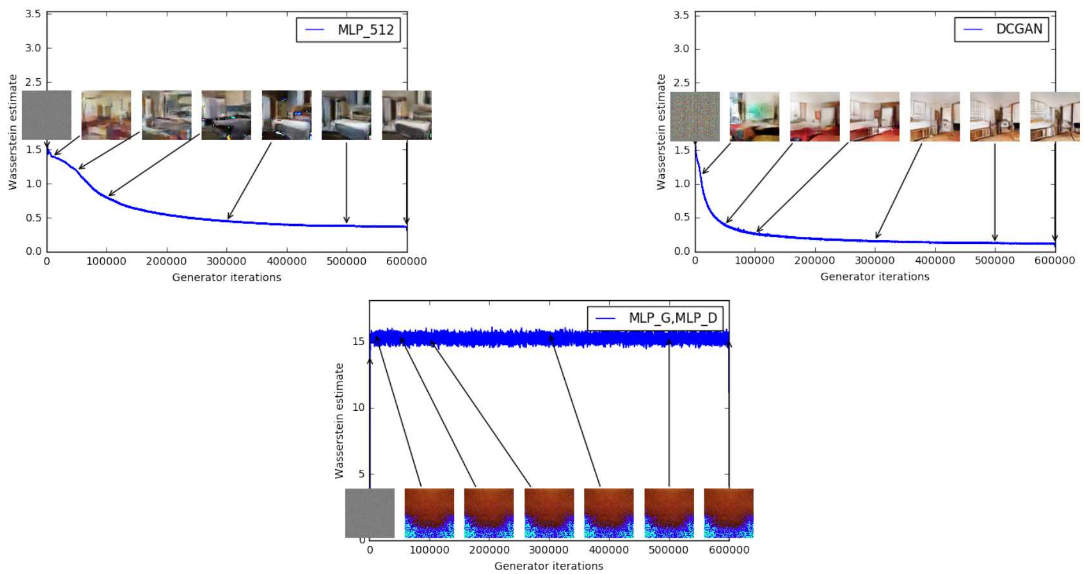

Figure 4: Training curves and samples at different stages of training. We can see a clear correlation between lower error and better sample quality. Upper left: the generator is an MLP with 4 hidden layers and 512 units at each layer. The loss decreases constistently as training progresses and sample quality increases. Upper right: the generator is a standard DCGAN. The loss decreases quickly and sample quality increases as well. In both upper plots the critic is a DCGAN without the sigmoid so losses can be subjected to comparison. Lower half: both the generator and the discriminator are MLPs with substantially high learning rates (so training failed). Loss is constant and samples are constant as well. The training curves were passed through a median filter for visualization purposes.

4. Empirical Results

We run experiments on image generation using our Wasserstein-GAN algorithm and show that there are significant practical benefits to using it over the formulation used in standard GANs. We claim two main benefits:

- a meaningful loss metric that correlates with the generator's convergence and sample quality
- improved stability of the optimization process

4.1. Mixtures of Gaussians

In (Metz et al., 2016) the authors presented a simple mixture of Gaussians experiments that served a very specific purpose. In this mixture, the mode collapse problem of GANs is easy to visualize, since a normal GAN would rotate between the different modes of the mixture, and fail to capture the whole distribution. In 2 we show how our WGAN algorithm approximately finds the correct distribution, without any mode collapse.

An interesting thing is that the WGAN first seems to learn to match the low-dimensional structure of the data (the approximate circle), before zooming in on the specific bumps of the true density. Similar to the Wasserstein distance, it looks like WGAN gives more importance to matching the low dimensional supports rather than the specific ratios between the densities.

4.2. Experimental Procedure for Image Generation

We run experiments on image generation. The target distribution to learn is the LSUN-Bedrooms dataset (Yu et al., 2015) – a collection of natural images of indoor bedrooms. Our baseline comparison is DCGAN (Radford et al., 2015), a GAN with a convolutional architecture trained with the standard GAN procedure using the $-\log D$ trick (Goodfellow et al., 2014). The generated samples are 3-channel images of 64x64 pixels in size. We use the hyper-parameters specified in Algorithm 1 for all of our experiments.

4.3. Meaningful loss metric

Because the WGAN algorithm attempts to train the critic f (lines 2–8 in Algorithm 1) relatively well before each generator update (line 10 in Algorithm 1), the loss function at this point is an estimate of the EM distance, up to constant factors related to the way we constrain the Lipschitz constant of f.

Our first experiment illustrates how this estimate correlates well with the quality of the generated samples. Besides the convolutional DCGAN architecture, we also ran experiments where we replace the generator or both the generator and the critic by 4-layer ReLU-MLP with 512 hidden units.

Figure 4 plots the evolution of the WGAN estimate (3) of the EM distance during WGAN training for all three architectures. The plots clearly show that these curves correlate

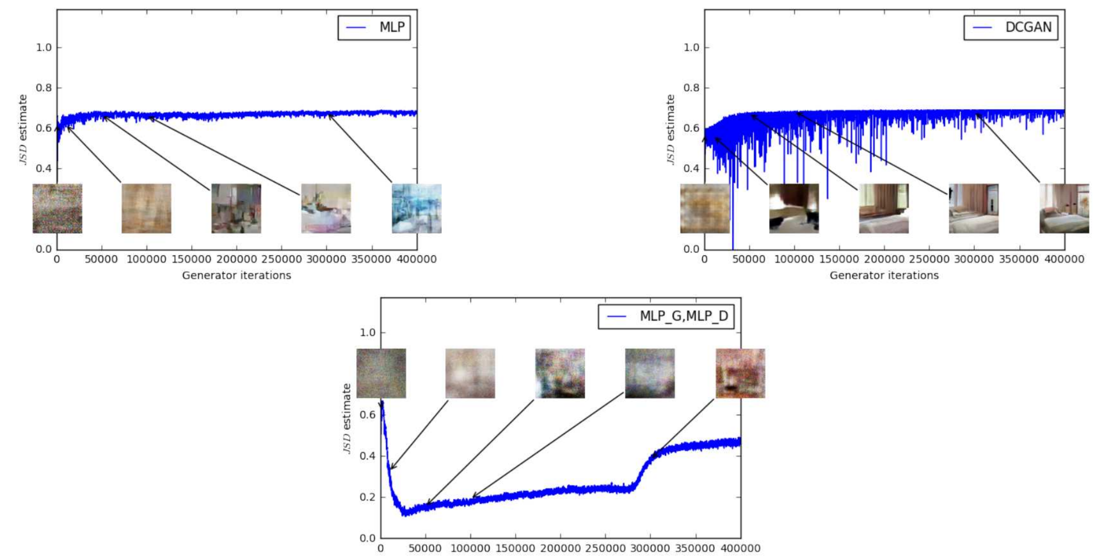

Figure 5: JS estimates for an MLP generator (upper left) and a DCGAN generator (upper right) trained with the standard GAN procedure. Both had a DCGAN discriminator. Both curves have increasing error. *Samples get better for the DCGAN but the JS estimate increases or stays constant, pointing towards no significant correlation between sample quality and loss. Bottom: MLP with both generator and discriminator. The curve goes up and down regardless of sample quality. All training curves were passed through the same median filter as in Figure 4.*

well with the visual quality of the generated samples.

To our knowledge, this is the first time in GAN literature that such a property is shown, where the loss of the GAN shows properties of convergence. This property is extremely useful when doing research in adversarial networks as one does not need to stare at the generated samples to figure out failure modes and to gain information on which models are doing better over others.

However, we do not claim that this is a new method to quantitatively evaluate generative models yet. The constant scaling factor that depends on the critic's architecture means it's hard to compare models with different critics. Even more, in practice the fact that the critic doesn't have infinite capacity makes it hard to know just how close to the EM distance our estimate really is. This being said, we have succesfully used the loss metric to validate our experiments repeatedly and without failure, and we see this as a huge improvement in training GANs which previously had no such facility.

In contrast, Figure 5 plots the evolution of the GAN estimate of the JS distance during GAN training. More precisely, during GAN training, the discriminator is trained to maximize

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

which is is a lower bound of $2JS(\mathbb{P}_r, \mathbb{P}_\theta) - 2\log 2$. In the figure, we plot the quantity $\frac{1}{2}L(D, g_\theta) + \log 2$, which is a lower bound of the JS distance.

This quantity clearly correlates poorly the sample quality. Note also that the JS estimate usually stays constant or goes up instead of going down. In fact it often remains very close to $\log 2 \approx 0.69$ which is the highest value taken by the JS distance. In other words, the JS distance saturates, the discriminator has zero loss, and the generated samples are in some cases meaningful (DCGAN generator, top right plot) and in other cases collapse to a single nonsensical image (Goodfellow et al., 2014). This last phenomenon has been theoretically explained in (Arjovsky & Bottou, 2017) and highlighted in (Metz et al., 2016).

When using the $-\log D$ trick (Goodfellow et al., 2014), the discriminator loss and the generator loss are different. Figure 9 in Appendix F reports the same plots for GAN training, but using the generator loss instead of the discriminator loss. This does not change the conclusions.

Finally, as a negative result, we report that WGAN training becomes unstable at times when one uses a momentum based optimizer such as Adam (Kingma & Ba, 2014) (with $\beta_1 > 0$) on the critic, or when one uses high learning rates. Since the loss for the critic is nonstationary, momentum based methods seemed to perform worse. We identified momentum as a potential cause because, as the loss blew up and samples got worse, the cosine between the Adam step and the gradient usually turned negative. The only places where this cosine was negative was in these situations of instability. We therefore switched to RMSProp (Tieleman & Hinton, 2012) which is known to perform well even on very nonstationary problems (Mnih et al., 2016).

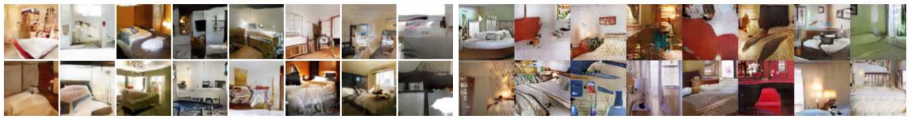

Figure 6: Algorithms trained with a DCGAN generator. Left: WGAN algorithm. Right: standard GAN formulation. Both algorithms produce high quality samples.

Figure 7: Algorithms trained with a generator without batch normalization and constant number of filters at every layer (as opposed to duplicating them every time as in (Radford et al., 2015)). Aside from taking out batch normalization, the number of parameters is therefore reduced by a bit more than an order of magnitude. Left: WGAN algorithm. Right: standard GAN formulation. As we can see the standard GAN failed to learn while the WGAN still was able to produce samples.

Figure 8: Algorithms trained with an MLP generator with 4 layers and 512 units with ReLU nonlinearities. The number of parameters is similar to that of a DCGAN, but it lacks a strong inductive bias for image generation. Left: WGAN algorithm. Right: standard GAN formulation. The WGAN method still was able to produce samples, lower quality than the DCGAN, and of higher quality than the MLP of the standard GAN. Note the significant degree of mode collapse in the GAN MLP.

4.4. Improved stability

One of the benefits of WGAN is that it allows us to train the critic till optimality. When the critic is trained to completion, it simply provides a loss to the generator that we can train as any other neural network. This tells us that we no longer need to balance generator and discriminator's capacity properly. The better the critic, the higher quality the gradients we use to train the generator.

We observe that WGANs are more robust than GANs when one varies the architectural choices for the generator in certain ways. We illustrate this by running experiments on three generator architectures: (1) a convolutional DCGAN generator, (2) a convolutional DCGAN generator without batch normalization and with a constant number of filters (the capacity of the generator is drastically smaller than that of the discriminator), and (3) a 4-layer ReLU-MLP with 512 hidden units. The last two are known to perform very poorly with GANs. We keep the convolutional DCGAN architecture for the WGAN critic or the GAN discriminator.

Figures 6, 7, and 8 show samples generated for these three architectures using both the WGAN and GAN algorithms. We refer the reader to Appendix H for full sheets of generated samples. Samples were not cherry-picked.

In no experiment did we see evidence of mode collapse for the WGAN algorithm.

5. Related Work

We refer the reader to Appendix D for the connections to the different integral probability metrics (Müller, 1997).

The recent work of (Montavon et al., 2016) has explored the use of Wasserstein distances in the context of learning for Restricted Boltzmann Machines for discrete spaces. Even though the motivations at a first glance might seem quite different, at the core of it both our works want to compare distributions in a way that leverages the geometry of the underlying space, which Wasserstein allows us to do.

Finally, the work of (Genevay et al., 2016) shows new algorithms for calculating Wasserstein distances between different distributions. We believe this direction is quite important, and perhaps could lead to new ways to evaluate generative models.

6. Conclusion

We introduced an algorithm that we deemed WGAN, an alternative to traditional GAN training. In this new model, we showed that we can improve the stability of learning, get rid of problems like mode collapse, and provide meaningful learning curves useful for debugging and hyperparameter searches. Furthermore, we showed that the corresponding optimization problem is sound, and provided extensive theoretical work highlighting the deep connections to other distances between distributions.

Acknowledgments

We would like to thank Mohamed Ishmael Belghazi, Emily Denton, Ian Goodfellow, Ishaan Gulrajani, Alex Lamb, David Lopez-Paz, Eric Martin, musyoku, Maxime Oquab, Aditya Ramesh, Ronan Riochet, Uri Shalit, Pablo Sprechmann, Arthur Szlam, Ruohan Wang, for helpful comments and advice.

References

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

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

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

Goodfellow, Ian J., 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 27*, pp. 2672–2680. Curran Associates, Inc., 2014.

Gretton, Arthur, Borgwardt, Karsten M., Rasch, Malte J., Schölkopf, Bernhard, and Smola, Alexander. A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773, 2012. ISSN 1532-4435.

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

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

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

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

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

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

Milgrom, Paul and Segal, Ilya. Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2):583–601, 2002. ISSN 1468-0262.

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

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

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

Neal, Radford M. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, April 2001. ISSN 0960-3174.

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

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

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

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

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

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

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

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

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

A. Why Wasserstein is indeed weak

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

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

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

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

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

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

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

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

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

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

We can consider

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

where $\Phi(\mathbb{P})(f) := \mathbb{E}_{x\sim\mathbb{P}}[f(x)]$ is a linear function over $C_b(\mathcal{X})$. The Riesz Representation theorem ((Kakutani, 1941), Theorem 10) tells us that Φ is an isometric immersion. This tells us that we can effectively consider Prob$(\mathcal{X})$ with the total variation distance as a subset of $C_b(\mathcal{X})^*$ with the norm distance. Thus, just to accentuate it one more time, the total variation over Prob$(\mathcal{X})$ is exactly the norm distance over $C_b(\mathcal{X})^*$.

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

Now, all dual spaces (such as $C_b(\mathcal{X})^*$ and thus Prob$(\mathcal{X})$) have a strong topology (induced by the norm), and a weak* topology. As the name suggests, the weak* topology is much weaker than the strong topology. In the case of Prob$(\mathcal{X})$, the strong topology is given by the total variation distance, and the weak* topology is given by the Wasserstein distance (among others) (Villani, 2009).

B. Assumption definitions

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

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

C. Proofs of things

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

By the definition of the Wasserstein distance, we have

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

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

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

Finally, we have that

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

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

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

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

By taking expectations and $z' = z$ we

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

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

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

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

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

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

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

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

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

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

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

finishing the proof □

Proof of Theorem 2.

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

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

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

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

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

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

Therefore,

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

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

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

We now can see that

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

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

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

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

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

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

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

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

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

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

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

2. This is a long known fact that W metrizes the weak* topology of $(C(\mathcal{X}), \| \cdot \|_\infty)$ on Prob$(\mathcal{X})$, and by definition this is the topology of convergence in distribution. A proof of this can be found (for example) in (Villani, 2009).

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

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

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

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

$\square$

Proof of Theorem 3. Let us define

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

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

Since $\mathcal{X}$ is compact, we know by the Kantorovich-Rubinstein duality (Villani, 2009) that there is an $f \in \mathcal{F}$ that attains the value

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

Let us define $X^*(\theta) = \{f \in \mathcal{F} : V(f, \theta) = W(\mathbb{P}_r, \mathbb{P}_\theta)\}$. By the above point we know then that $X^*(\theta)$ is non-empty. We know by a simple envelope theorem ((Milgrom & Segal, 2002), Theorem 1) that

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

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

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

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

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

$$-\nabla_\theta \mathbb{E}_{z \sim p(z)}[f(g_\theta(z))] = -\mathbb{E}_{z \sim p(z)}[\nabla_\theta f(g_\theta(z))] \tag{4}$$

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

Since $f \in \mathcal{F}$, we know that it is 1-Lipschitz. Furthermore, $g_\theta(z)$ is locally Lipschitz as a function of (θ, z). Therefore, $f(g_\theta(z))$ is locally Lipschitz on (θ, z) with constants $L(\theta, z)$ (the same ones as g). By Radamacher's Theorem, $f(g_\theta(z))$ has to be differentiable almost everywhere for (θ, z) jointly. Rewriting this, the set $A = \{(\theta, z) : f \circ g \text{ is not differentiable}\}$ has measure 0. By Fubini's Theorem, this implies that for almost every θ the section $A_\theta = \{z : (\theta, z) \in A\}$ has measure 0. Let's now fix a θ_0 such that the measure of A_{θ_0} is null (**such as when the right hand side of equation (4) is well defined**). For this θ_0 we have $\nabla_\theta f(g_\theta(z))|_{\theta_0}$ is well-defined for almost any z, and since $p(z)$ has a density, it is defined $p(z)$-a.e. By assumption 1 we know that

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

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

$$\frac{\mathbb{E}_{z \sim p(z)}[f(g_\theta(z))] - \mathbb{E}_{z \sim p(z)}[f(g_{\theta_0}(z))] - \langle (\theta - \theta_0), \mathbb{E}_{z \sim p(z)}[\nabla_\theta f(g_\theta(z))|_{\theta_0}] \rangle}{\|\theta - \theta_0\|} \tag{5}$$

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

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

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

and since $\mathbb{E}_{z \sim p(z)}[2L(\theta_0, z)] < +\infty$ by assumption 1, we get by dominated convergence that Equation 5 converges to 0 as $\theta \to \theta_0$ so

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

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

D. Related Work

There's been a number of works on the so called Integral Probability Metrics (IPMs) (Müller, 1997). Given $\mathcal{F}$ a set of functions from $\mathcal{X}$ to $\mathbb{R}$, we can define

$$d_{\mathcal{F}}(\mathbb{P}_r, \mathbb{P}_\theta) = \sup_{f \in \mathcal{F}} \mathbb{E}_{x \sim \mathbb{P}_r}[f(x)] - \mathbb{E}_{x \sim \mathbb{P}_\theta}[f(x)] \tag{6}$$

as an integral probability metric associated with the function class $\mathcal{F}$. It is easily verified that if for every $f \in \mathcal{F}$ we have $-f \in \mathcal{F}$ (such as all examples we'll consider), then $d_{\mathcal{F}}$ is nonnegative, satisfies the triangular inequality, and is symmetric. Thus, $d_{\mathcal{F}}$ is a pseudometric over Prob$(\mathcal{X})$.

While IPMs might seem to share a similar formula, as we will see different classes of functions can yield to radically different metrics.

- By the Kantorovich-Rubinstein duality (Villani, 2009), we know that $W(\mathbb{P}_r, \mathbb{P}_\theta) = d_{\mathcal{F}}(\mathbb{P}_r, \mathbb{P}_\theta)$ when $\mathcal{F}$ is the set of 1-Lipschitz functions. Furthermore, if $\mathcal{F}$ is the set of K-Lipschitz functions, we get $K \cdot W(\mathbb{P}_r, \mathbb{P}_\theta) = d_{\mathcal{F}}(\mathbb{P}_r, \mathbb{P}_\theta)$.

- When $\mathcal{F}$ is the set of all continuous functions bounded between -1 and 1, we retrieve $d_{\mathcal{F}}(\mathbb{P}_r, \mathbb{P}_\theta) = \delta(\mathbb{P}_r, \mathbb{P}_\theta)$ the total variation distance (Müller, 1997). This already tells us that going from 1-Lipschitz to 1-Bounded functions drastically changes the topology of the space, and the regularity of $d_{\mathcal{F}}(\mathbb{P}_r, \mathbb{P}_\theta)$ as a loss function (as by Theorems 1 and 2).

- Energy-based GANs (EBGANs) (Zhao et al., 2016) can be thought of as the generative approach to the total variation distance. This connection is stated and proven in depth in Appendix E. At the core of the connection is that the discriminator will play the role of f maximizing equation (6) while its only restriction is being between 0 and m for some constant m. This will yield the same behaviour as being restricted to be between -1 and 1 up to a constant scaling factor irrelevant to optimization. Thus, when the discriminator approaches optimality the cost for the generator will aproximate the total variation distance $\delta(\mathbb{P}_r, \mathbb{P}_\theta)$.

  Since the total variation distance displays the same regularity as the JS, it can be seen that EBGANs will suffer from the same problems of classical GANs regarding not being able to train the discriminator till optimality and thus limiting itself to very imperfect gradients.

- Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) is a specific case of integral probability metrics when $\mathcal{F} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq 1\}$ for $\mathcal{H}$ some Reproducing Kernel Hilbert Space (RKHS) associated with a given kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. As proved on (Gretton et al., 2012) we know that MMD is a proper metric and not only a pseudometric when the kernel is universal. In the specific case where $\mathcal{H} = L^2(\mathcal{X}, m)$ for m the normalized Lebesgue measure on $\mathcal{X}$, we know that $\{f \in C_b(\mathcal{X}), \|f\|_\infty \leq 1\}$ will be contained in $\mathcal{F}$, and therefore $\delta(\mathbb{P}_r, \mathbb{P}_\theta) \leq d_{\mathcal{F}}(\mathbb{P}_r, \mathbb{P}_\theta)$, so the regularity of the MMD distance as a loss function will be at least as bad as the one of the total variation. Nevertheless this is a very extreme case, since we would need a very powerful kernel to approximate the whole L^2. However, even Gaussian kernels are able to detect tiny noise patterns as recently evidenced by (Sutherland et al., 2017). This points to the fact that especially with low bandwidth kernels, the distance might be close to a saturating regime similar as with total variation or the JS. This obviously doesn't need to be the case for every kernel, and figuring out how and which different MMDs are closer to Wasserstein or total variation distances is an interesting topic of research.

  The great aspect of MMD is that via the kernel trick there is no need to train a separate network to maximize equation (6) for the ball of a RKHS. However, this has the disadvantage that evaluating the MMD distance has computational cost that grows quadratically with the amount of samples used to estimate the expectations in (6). This last point makes MMD have limited scalability, and is sometimes inapplicable to many real life applications because of it. There are estimates with linear computational cost for the MMD (Gretton et al., 2012) which in a lot of cases makes MMD very useful, but they also have worse sample complexity.

- Generative Moment Matching Networks (GMMNs) (Li et al., 2015; Dziugaite et al., 2015) are the generative counterpart of MMD. By backproping through the kernelized formula for equation (6), they directly optimize $d_{MMD}(\mathbb{P}_r, \mathbb{P}_\theta)$ (the IPM when $\mathcal{F}$ is as in the previous item). As mentioned, this has the advantage of not requiring a separate network to approximately maximize equation (6). However, GMMNs have enjoyed limited applicability. Partial explanations for their unsuccess are the quadratic cost as a function of the number of samples and vanishing gradients for low-bandwidth kernels. Furthermore, it may be possible that some kernels used in practice are unsuitable for capturing very complex distances in high dimensional sample spaces such as natural images. This is properly justified by the fact that (Ramdas et al., 2014) shows that for the typical Gaussian MMD test to be reliable (as in it's power as a statistical test approaching 1), we need the number of samples to grow linearly with the number of dimensions. Since the MMD computational cost grows quadratically with the number of samples in the batch used to estimate equation (6), this makes the cost of having a reliable estimator grow quadratically with the number of dimensions, which makes it very inapplicable for high dimensional problems. Indeed, for something as standard as $64x64$ images, we would need minibatches of size at least 4096 (without taking into account the constants in the bounds of (Ramdas et al., 2014) which would make this number substantially larger) and a total cost per iteration of 4096^2, over 5 orders of magnitude more than a GAN iteration when using the standard batch size of 64.

That being said, these numbers can be a bit unfair to the MMD, in the sense that we are comparing empirical sample complexity of GANs with the theoretical sample complexity of MMDs, which tends to be worse. However, in the original GMMN paper (Li et al., 2015) they indeed used a minibatch of size 1000, much larger than the standard 32 or 64 (even when this incurred in quadratic computational cost). While estimates that have linear computational cost as a function of the number of samples exist (Gretton et al., 2012), they have worse sample complexity, and to the best of our knowledge they haven't been yet applied in a generative context such as in GMMNs.

On another great line of research, the recent work of (Montavon et al., 2016) has explored the use of Wasserstein distances in the context of learning for Restricted Boltzmann Machines for discrete spaces. The motivations at a first glance might seem quite different, since the manifold setting is restricted to continuous spaces and in finite discrete spaces the weak and strong topologies (the ones of W and JS respectively) coincide. However, in the end there is more in commmon than not about our motivations. We both want to compare distributions in a way that leverages the geometry of the underlying space, and Wasserstein allows us to do exactly that.

Finally, the work of (Genevay et al., 2016) shows new algorithms for calculating Wasserstein distances between different distributions. We believe this direction is quite important, and perhaps could lead to new ways to evaluate generative models.

E. Energy-based GANs optimize total variation

In this appendix we show that under an optimal discriminator, energy-based GANs (EBGANs) (Zhao et al., 2016) optimize the total variation distance between the real and generated distributions.

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

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

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

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

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

We say that a measurable function $D^* : \mathcal{X} \to [0, +\infty)$ is optimal for g_θ (or $\mathbb{P}_\theta$) if $L_D(D^*, g_\theta) \leq L_D(D, g_\theta)$ for all other measurable functions D. We show that such a discriminator always exists for any two distributions $\mathbb{P}_r$ and $\mathbb{P}_\theta$, and that under such a discriminator, $L_G(D^*, g_\theta)$ is proportional to $\delta(\mathbb{P}_r, \mathbb{P}_\theta)$. As a simple corollary, we get the fact that $L_G(D^*, g_\theta)$ attains its minimum value if and only if $\delta(\mathbb{P}_r, \mathbb{P}_\theta)$ is at its minimum value, which is 0, and $\mathbb{P}_r = \mathbb{P}_\theta$ (Theorems 1-2 of (Zhao et al., 2016)).

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

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

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

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

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

Therefore, we know that

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

The interesting part is that

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

and there is an $f^* : \mathcal{X} \to [-1, 1]$ such that $\mathbb{E}_{x\sim\mathbb{P}_r}[f^*(x)] - \mathbb{E}_{x\sim\mathbb{P}_\theta}[f^*(x)] = -\delta(\mathbb{P}_r, \mathbb{P}_\theta)$. This is a long known fact, found for example in (Villani, 2009), but we prove it later for completeness. In that case, we define $D^*(x) = \frac{m}{2}f^*(x) + \frac{m}{2}$. We then have $0 \leq D(x) \leq m$ and

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

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

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

concluding the proof.

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

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

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

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

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

F. Generator's cost during normal GAN training

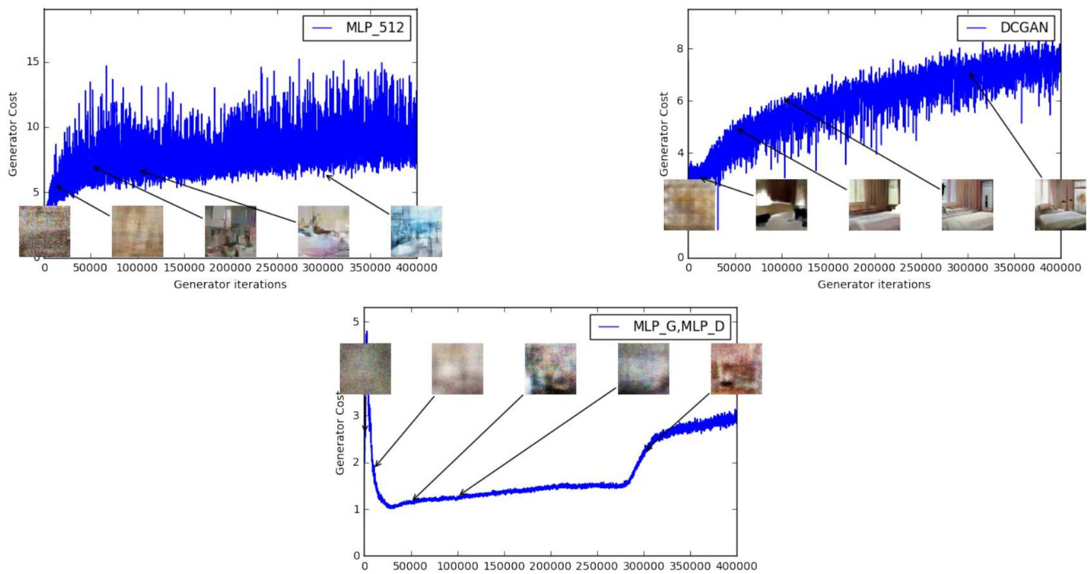

Figure 9: Cost of the generator during normal GAN training, for an MLP generator (upper left) and a DCGAN generator (upper right). Both had a DCGAN discriminator. ***Both curves have increasing error.*** *Samples get better for the DCGAN but the cost of the generator increases, pointing towards no significant correlation between sample quality and loss. Bottom: MLP with both generator and discriminator. The curve goes up and down regardless of sample quality. All training curves were passed through the same median filter as in Figure 4.*

G. Further comments on clipping and batch normalization

In this appendix we provide further informal insights into the behaviour of weight clipping and batch normalization in the context of GANs and WGANs.

G.1. Weight clipping

One may wander what would happen if one were to use weight clipping in a standard GAN. Disregarding the use of the cross-entropy vs difference loss, the central difference would be the use of a sigmoid in the end of the discriminator. This brings into place the use of the Dudley metric:

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

where

$$\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R}, f \text{ continuous and } \|f\|_\infty + \|f\|_L \leq 1\}$$

This is similar to the class of 1-Lipschitz functions, only we restrict how high the values of f can be. This metric is easily shown to be equivalent to the one with

$$\mathcal{F}'' = \{f : \mathcal{X} \to \mathbb{R}, f \text{ continuous and } \|f\|_\infty \leq 1, \|f\|_L \leq K\}$$

or the one with

$$\mathcal{F}' = \{f : \mathcal{X} \to \mathbb{R}, f \text{ continuous and } 0 \leq f \leq 1, \|f\|_L \leq K\} \tag{8}$$

This last family is essentially the family of functions we would achieve by adding a sigmoid to the critic of the WGAN, moving us closer to the standard GAN realm. An easy and very interesting result is that $d_{\mathcal{F}}$ and $d_{\mathcal{F}'}$ have the same topology as the Wasserstein distance (Villani, 2009), which hints that adding clipping to a GAN lands us closer to a WGAN than the standard GAN (since the cost function between distributions has the exact same regularity as that of a WGAN, and drastically different from a normal GAN, see Theorems 1, 2).

That being said, while the topology of Wasserstein and the Dudley metric is the same, some things are not. There are many distances that yield the weak topology, but Wasserstein has a number of differences against the rest. These are perfectly explained in pages 110 and 111 of (Villani, 2009), but we highlight the main idea here. At the core of it, Wasserstein is better at representing long distances between samples. The saturation behaviour of the sigmoid, or what happens when K in (8) is large, shows that if the real samples are far away from the fake ones, f can saturate at 1 in the real and 0 in the fake constantly, providing no usable gradient. Thus, Dudley and Wasserstein have the same behaviour in *close* samples, but Wasserstein avoids saturations and provides gradients even when samples are far away. However, if K (i.e. the clipping) is small enough (such as the 0.01 we use in practice), the saturating regime will never enter in place, so Dudley and Wasserstein will behave in the same way.

To conclude, if the clipping is small enough, the network is quite literally a WGAN, and if it's large it will saturate and fail to take into account information between samples that are far away (much like a normal GAN when the discriminator is trained till optimum).

As to the similarities between the difference vs cross-entropy on the loss of the discriminator or critic: if the supports of $\mathbb{P}_r$ and $\mathbb{P}_\theta$ are essentially disjoint (which was shown to happen in (Arjovsky & Bottou, 2017) in the usual setting of low dimensional supports), with both cost functions the f will simply be trained to attain the maximum value possible in the real and the minimum possible in the fake, without surpassing the Lipschitz constraint. Therefore, CE and the difference might behave more similarly than we expect in the typical 'learning low dimensional distributions' setting, provided we have a strong Lipschitz constraint.

G.2. Batch normalization

It is not clear that batch normalization (BN) is a Lipschitz function with a constant independent of the parameters, since we are dividing by the standard deviation, which depends on the inputs and the parameters. In this subsection we explain why BN still behaves in a Lipschitz way, and fits in with the theoretical support of our method.

Let x be the input of a BN layer. If there is a positive $c \in \mathbb{R}$ for which $V(x)^{1/2} > c$ (the variance is uniformly bounded below during training), then this $c \in \mathbb{R}$ becomes a Lipschitz constant on the BN layer that's independent of the model parameters, as we wanted. For $V(x)$ to not be bounded below as training progresses, it has to go arbitrarily close to 0. In this case, x has to converge (up to taking a subsequence) to it's mean, so the term $x - \mathbb{E}[x]$ in the numerator of batchnorm will go to 0, and therefore $\frac{x-\mathbb{E}[x]}{V[x]^{1/2}+\epsilon}$ comes the constant 0 (which is obviously 1-Lipschitz) due to the ϵ in the division. This will also further render the activation x inactive, which the network has no incentive to do.

While this argument is handwavy, one can formalize it and prove very simple bounds that depend only on ϵ. By increasing ϵ one can enforce a stronger Lipschitz constant, and we could have for example clamped the denominator of the BN to attain a value large enough. However, in practice in all our runs the variance never surpassed low thresholds, and this clamping of the BN division was simply never set into effect. Thus, we empirically never saw a break in the Lipschitness of our BN layers.

H. Sheets of samples

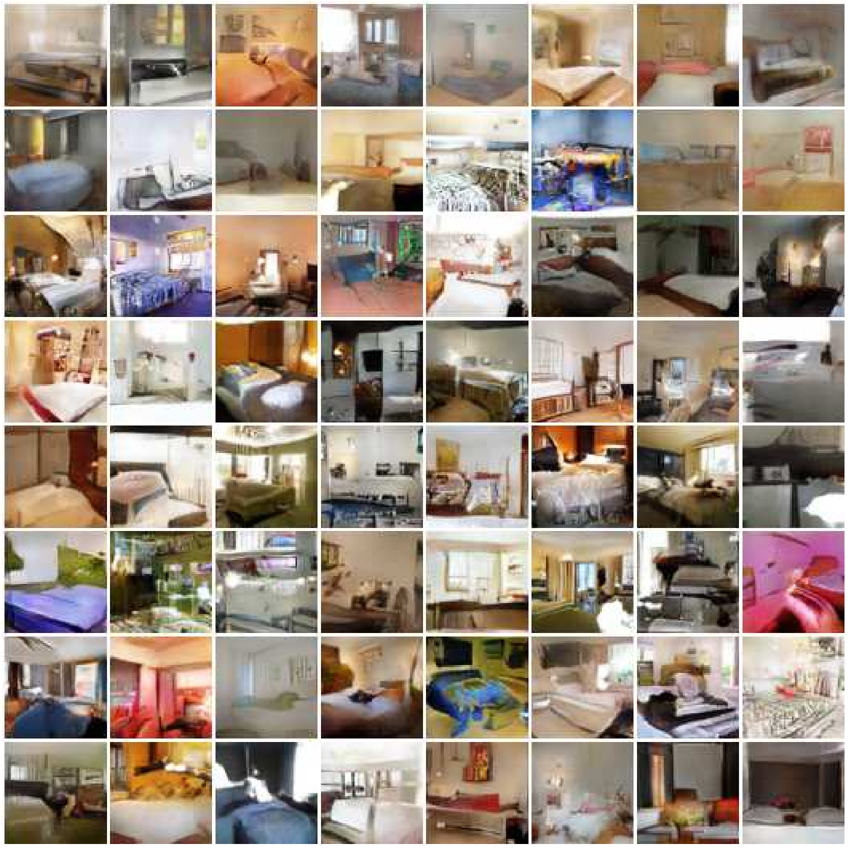

Figure 10: WGAN algorithm: generator and critic are DCGANs.

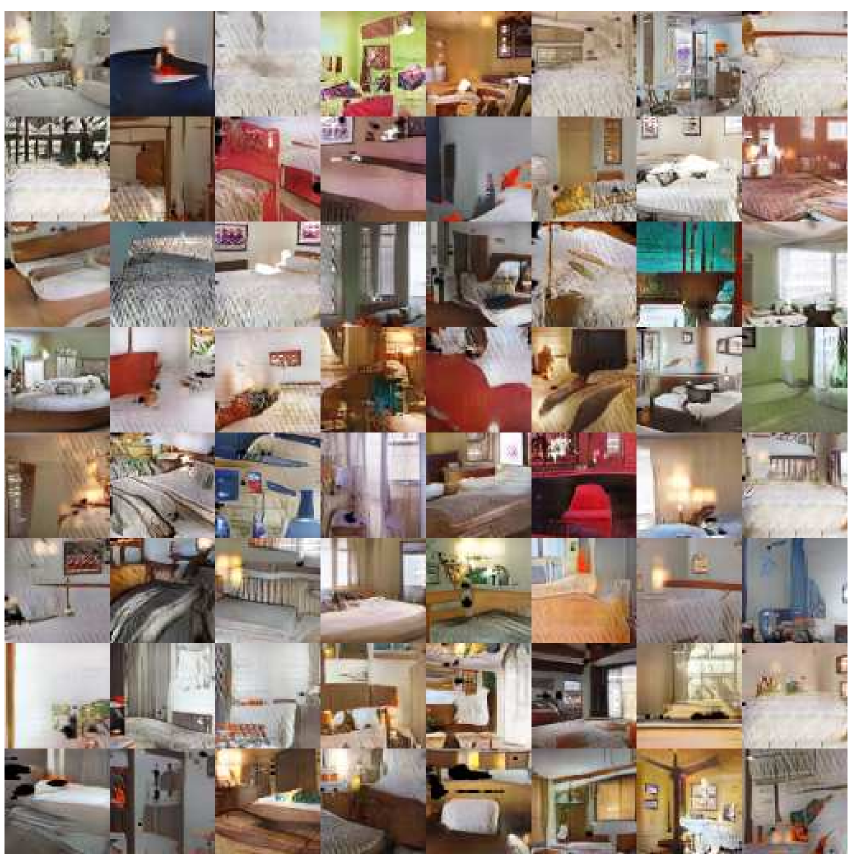

Figure 11: Standard GAN procedure: generator and discriminator are DCGANs.

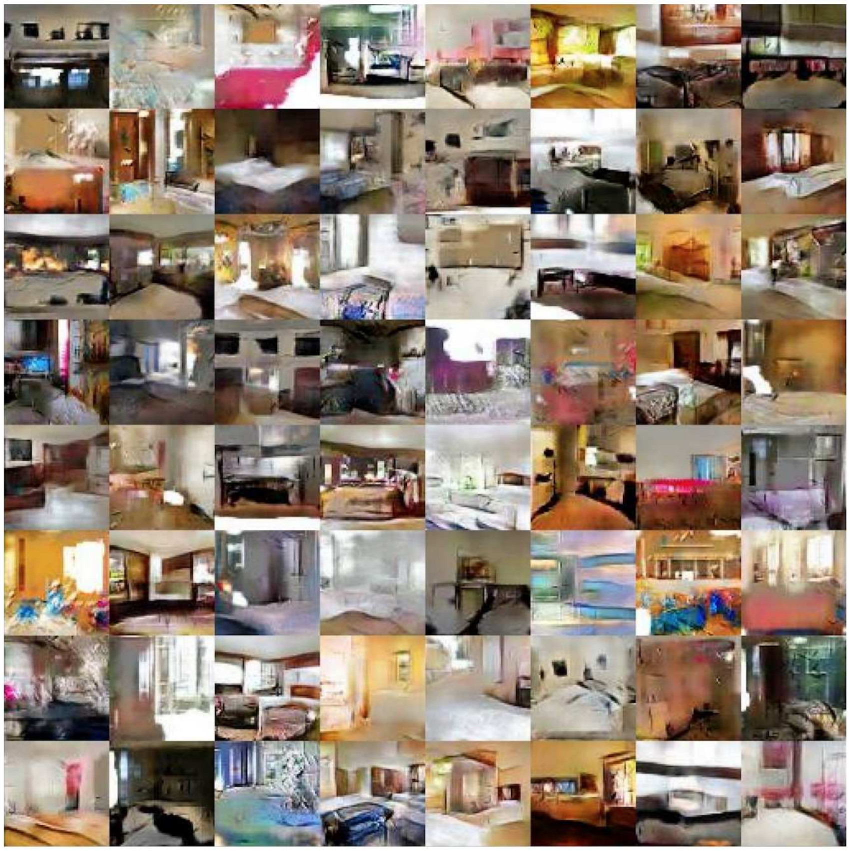

Figure 12: WGAN algorithm: generator is a DCGAN without batchnorm and constant filter size. Critic is a DCGAN.

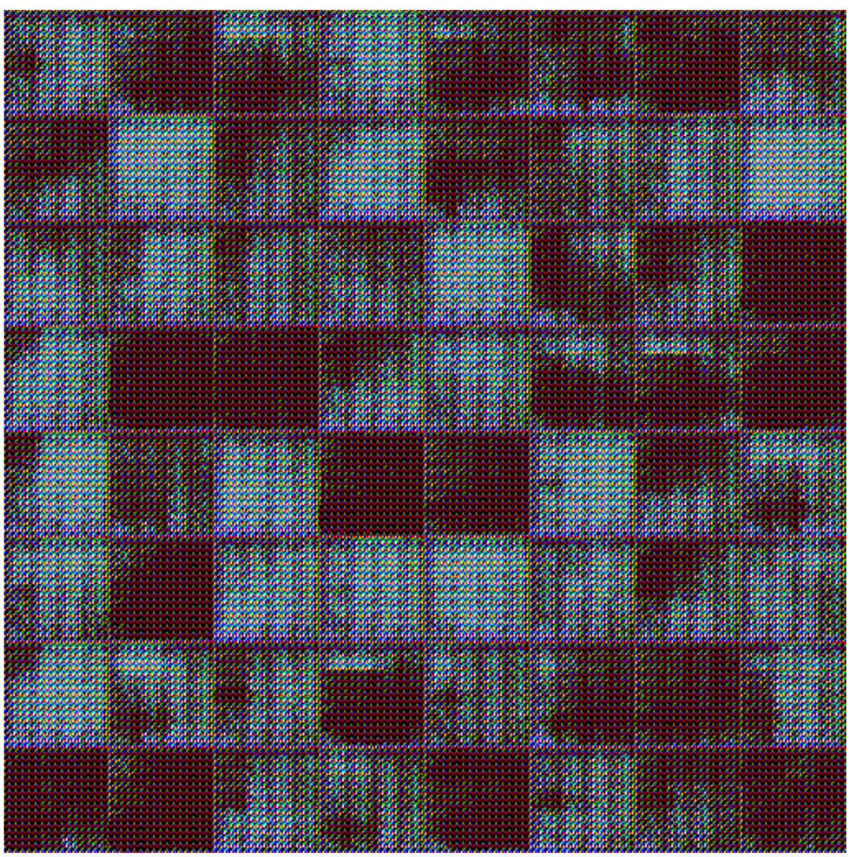

Figure 13: Standard GAN procedure: generator is a DCGAN without batchnorm and constant filter size. Discriminator is a DCGAN.

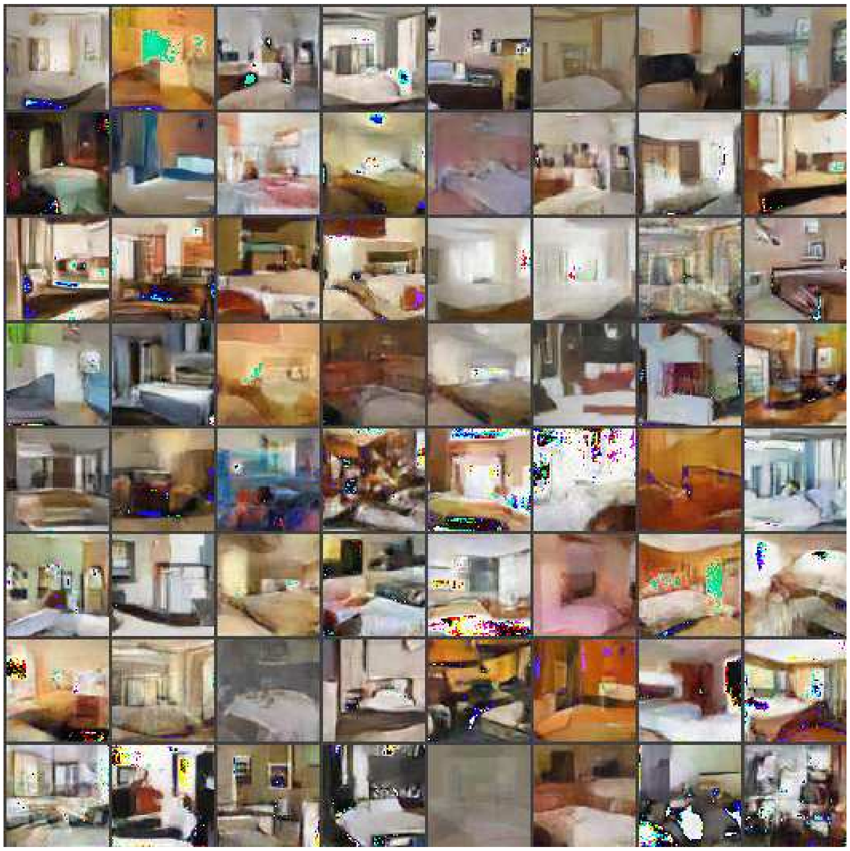

Figure 14: WGAN algorithm: generator is an MLP with 4 hidden layers of 512 units, critic is a DCGAN.

Figure 15: Standard GAN procedure: generator is an MLP with 4 hidden layers of 512 units, discriminator is a DCGAN.