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 $\text{Prob}(\mathcal{X})$ 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 $\text{Prob}(\mathcal{X})$ (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 $\text{Prob}(\mathcal{X})$ 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 $\text{Prob}(\mathcal{X})$ 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 $\text{Prob}(\mathcal{X})$) 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 $\text{Prob}(\mathcal{X})$, 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 $g_\theta(z) \rightarrow_{\theta\rightarrow\theta'} g_{\theta'}(z)$, so $\|g_\theta - g_{\theta'}\| \rightarrow 0$ 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

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

Finally, we have that

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

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

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

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$(\mathcal{X})$.

□

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. □

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 $\mathbb{E}_{x\sim\mathbb{P}_r}[f(x)] - \mathbb{E}_{x\sim\mathbb{P}_\theta}[f(x)] \geq -\delta(\mathbb{P}_r, \mathbb{P}_\theta)$. □

F. Generator's cost during normal GAN training

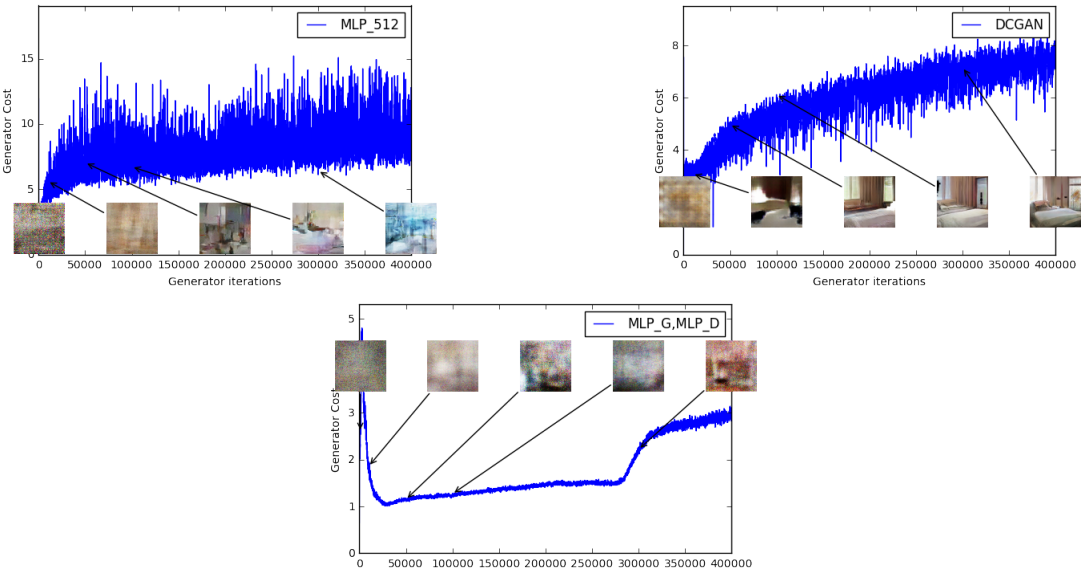

Figure 9: Cost of the generator during normal GAN training, for an MLP generator (upper left) and a DCGAN generator (upper right). Both had a DCGAN discriminator. ***Both curves have increasing error****. Samples get better for the DCGAN but the cost of the generator increases, pointing towards no significant 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

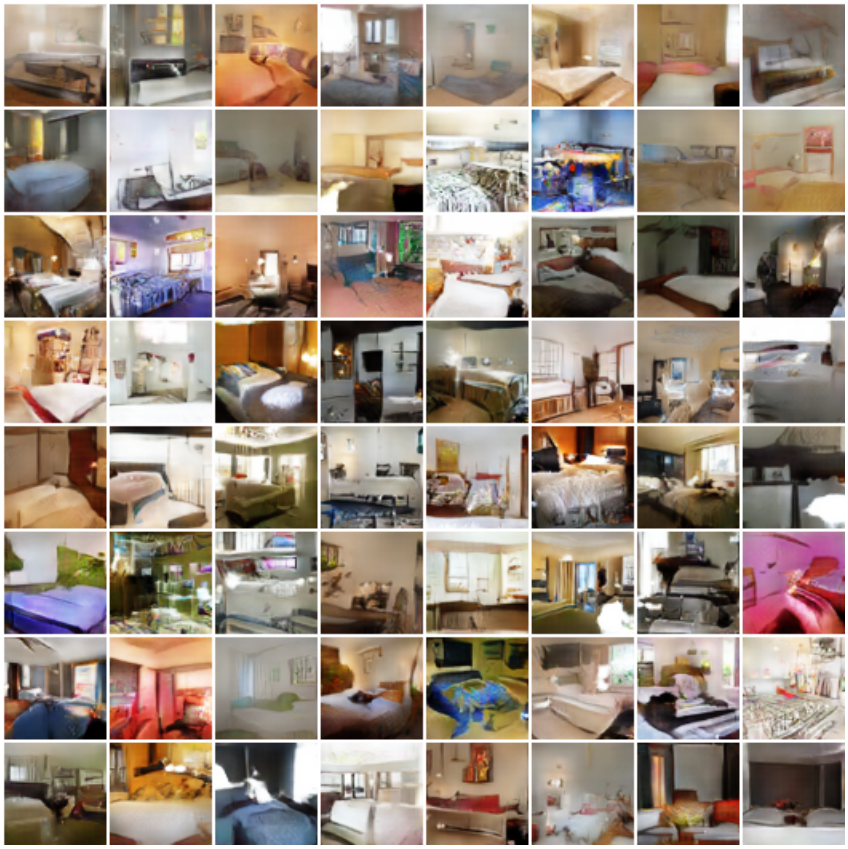

Figure 10: WGAN algorithm: generator and critic are DCGANs.

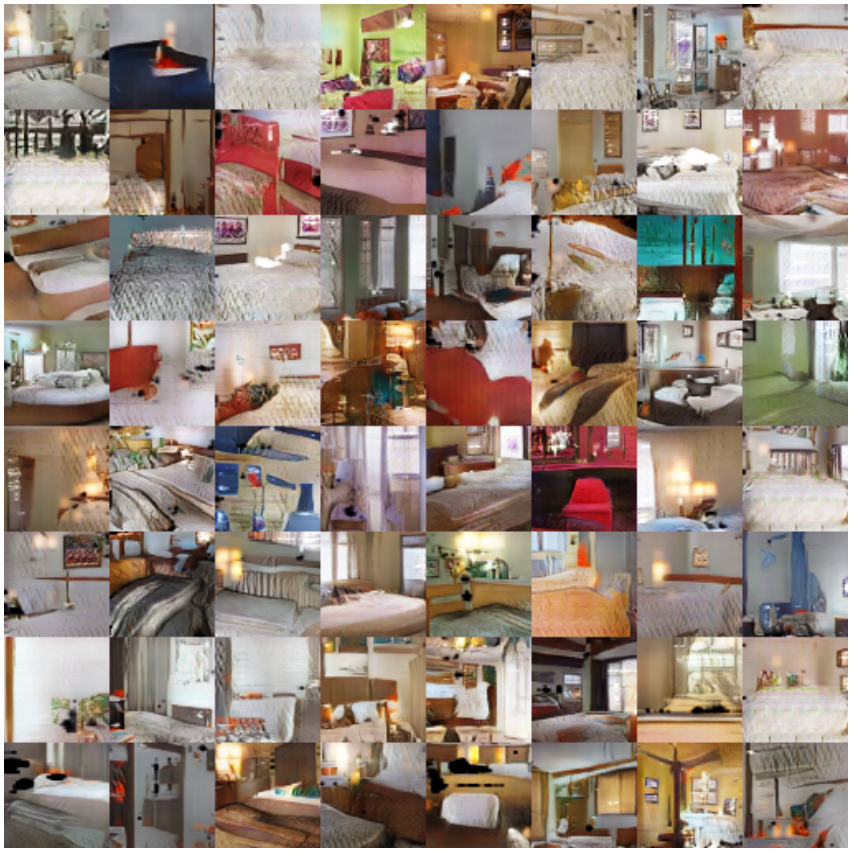

Figure 11: Standard GAN procedure: generator and discriminator are DCGANs.

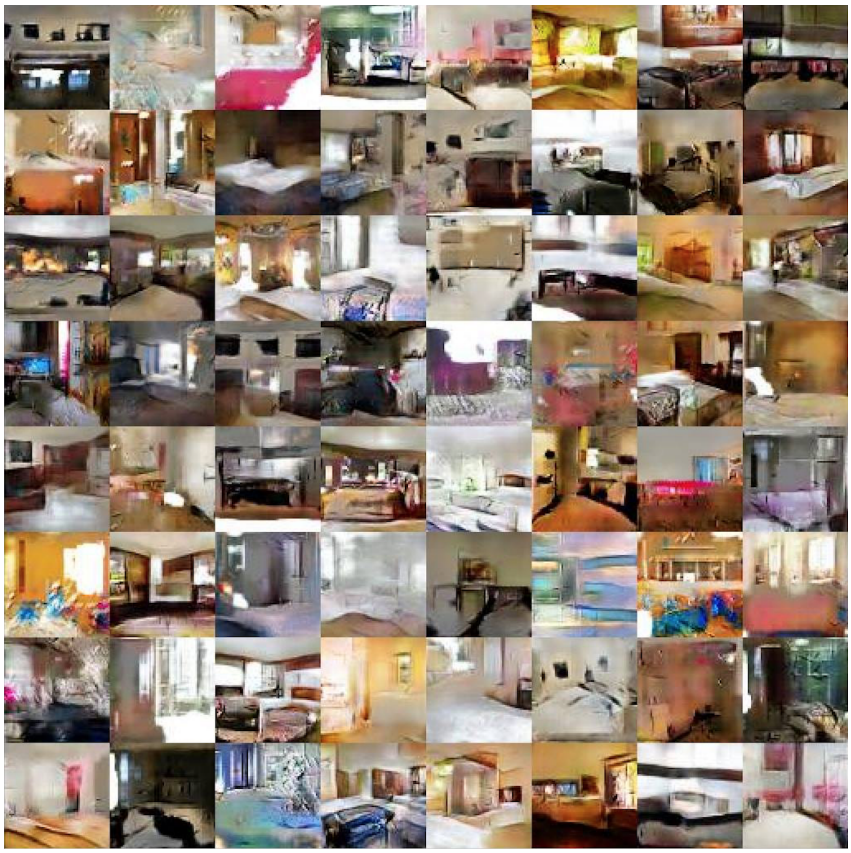

Figure 12: WGAN algorithm: generator is a DCGAN without batchnorm and constant filter size. Critic is a DCGAN.

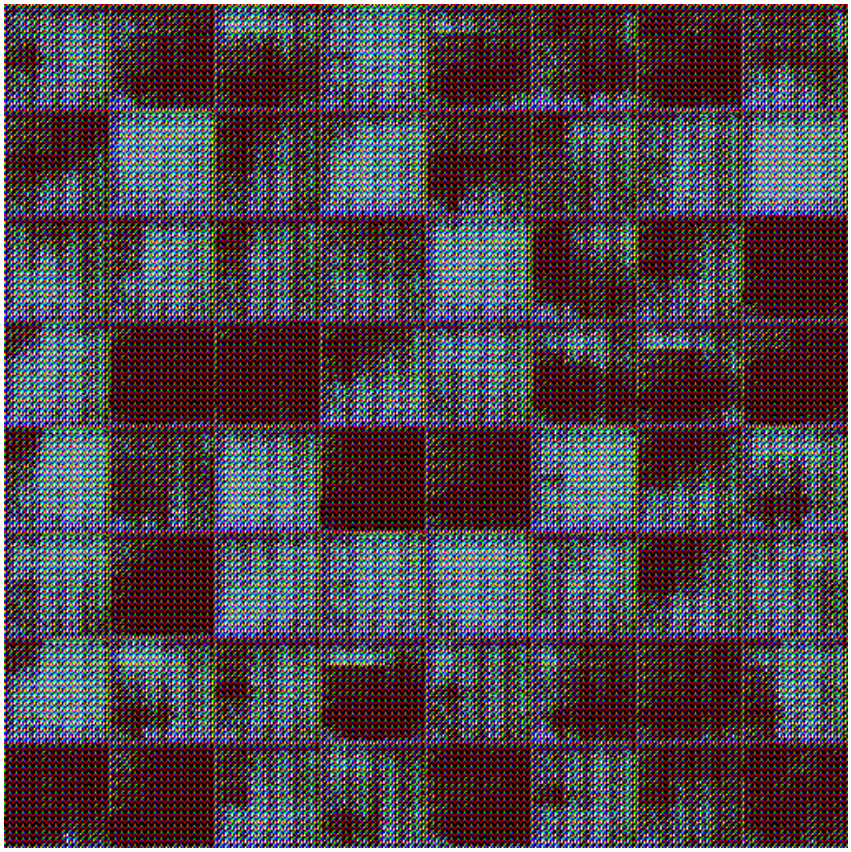

Figure 13: Standard GAN procedure: generator is a DCGAN without batchnorm and constant filter size. Discriminator is a DCGAN.

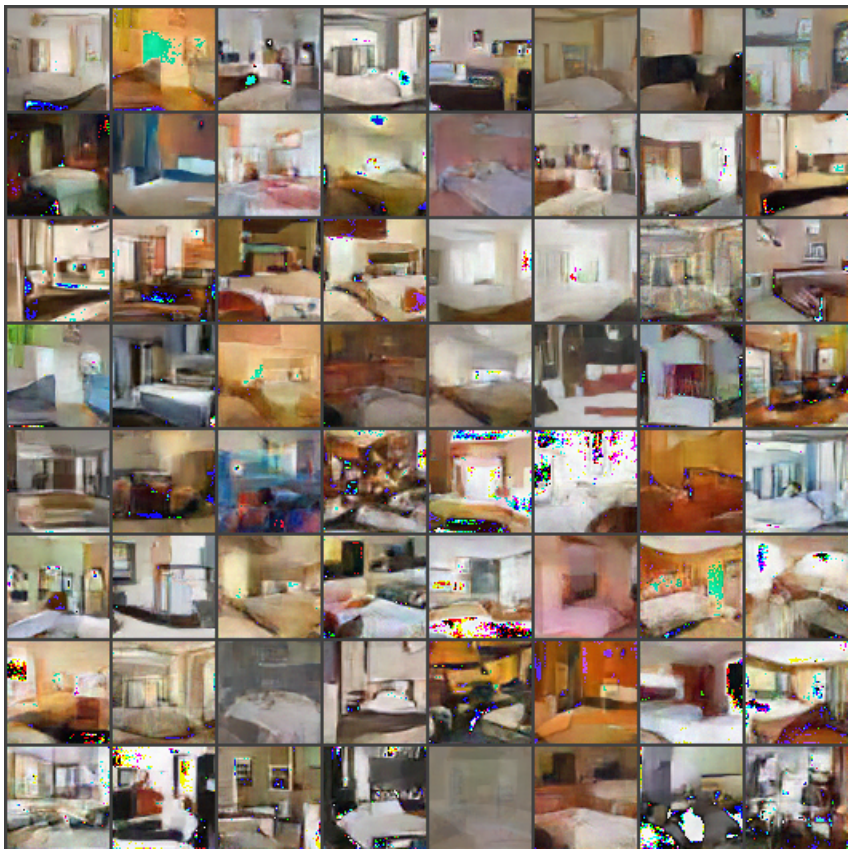

Figure 14: WGAN algorithm: generator is an MLP with 4 hidden layers of 512 units, critic is a DCGAN.

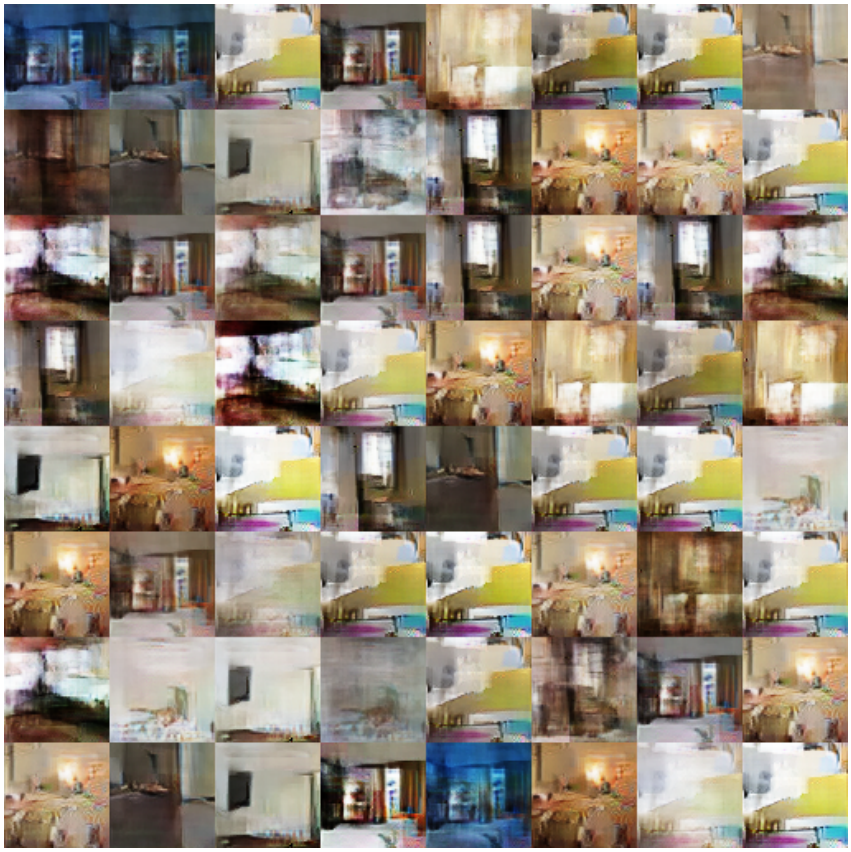

Figure 15: Standard GAN procedure: generator is an MLP with 4 hidden layers of 512 units, discriminator is a DCGAN.