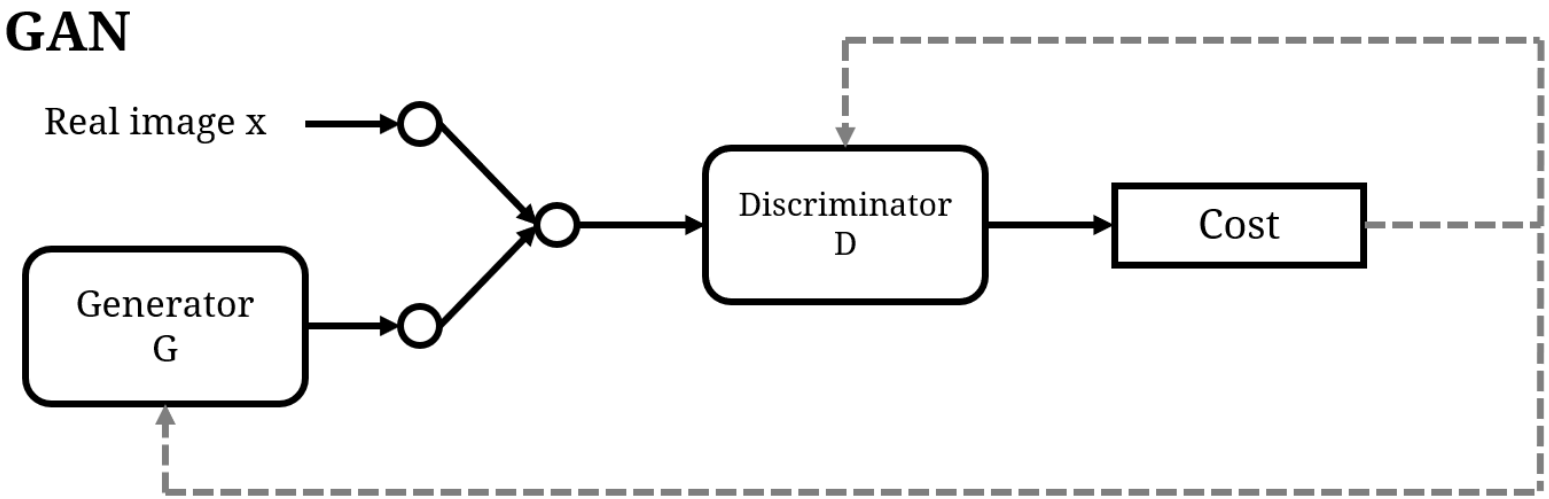
Wasserstein GAN

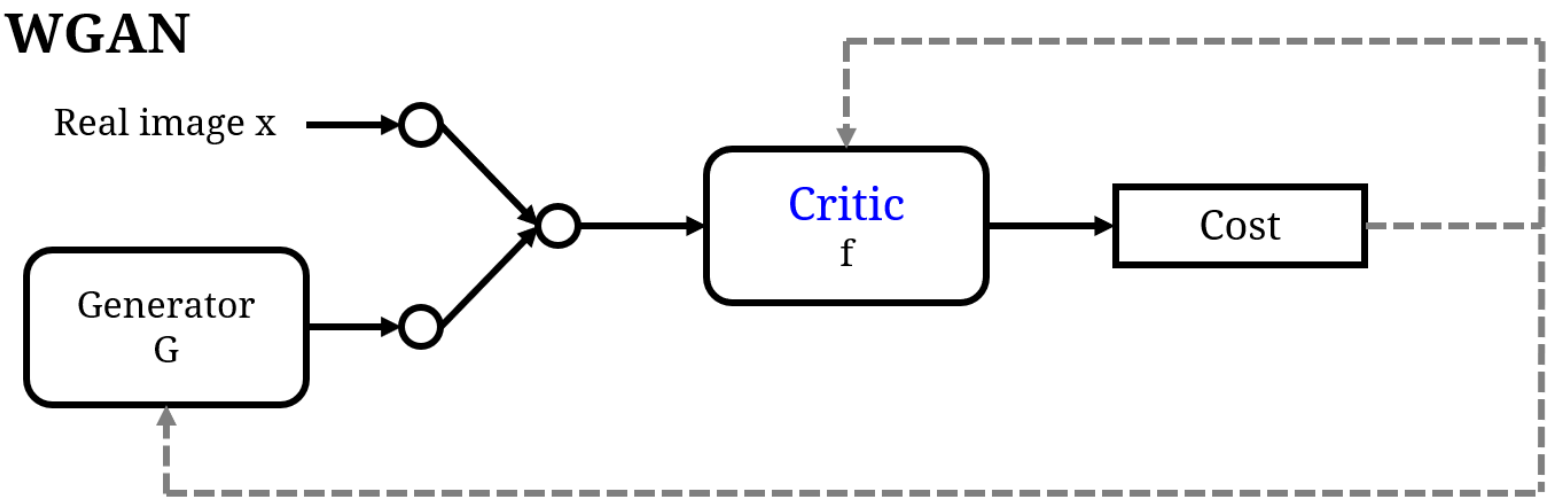
신준호, 박태준

공간통계연구실

2022년 7월 5일

Topic


GAN
Real image x
Generator
G
Discriminator
D
Cost


WGAN
Real image x
Generator
G
Critic
f
Cost

GAN의 문제점

- Minimizing the GAN objective function with an optimal discriminator, D^*, is equivalent to minimizing the JS-divergence;

$$\min_G V(D^*, G) = 2JS(\mathbb{P}_r, \mathbb{P}_\theta) - 2\log 2$$

- **Claim:** Divergence, which is not continuous with the generator's parameter, leads to difficulty in learning.

GAN의 문제점

Training GANs is hard for theoretical reasons with the GAN cost functions.[1]

- When $\mathbb{P}_r$ and $\mathbb{P}_\theta$ lie on low dimensional manifolds, there's always a perfect discriminator that can be trained well.
- It provides no usable gradients.
  ($\nabla D^*(x)$ will be 0 for almost everywhere.)
  - Gradient vanishing:

    $$\nabla_{\theta_g} \log \Big(1 - D(G(z^{(i)}))\Big) \to 0$$

    under optimal discriminator. (D is close to D^*)
  - Mode collapse:

    $$-\nabla_{\theta_g} \log D(G(z^{(i)}))$$

    unstable with large variance of gradients.

[1] Towards principled methods for training Generative Adversarial Networks, Arjovsky et al 2017

GAN의 문제점

Discriminator vs Critic

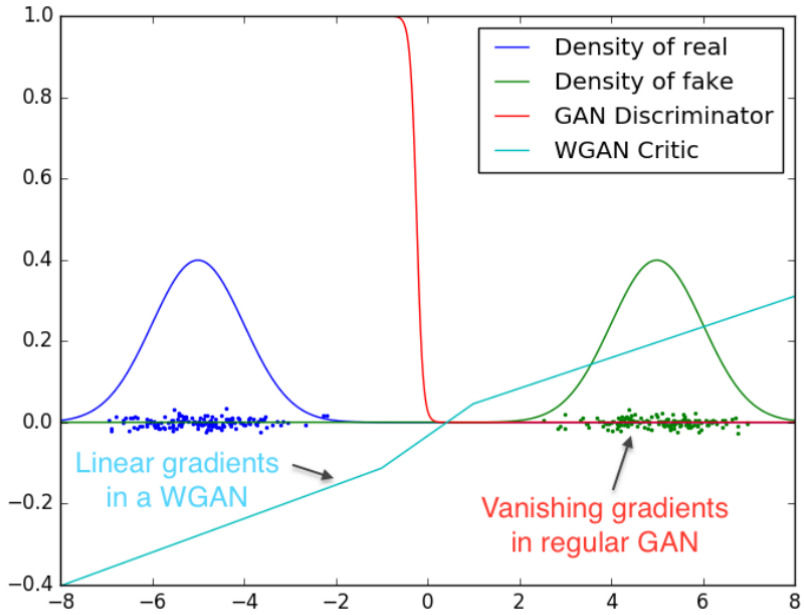


Figure 2: Optimal discriminator and critic when learning to differentiate two Gaussians. As we can see, the discriminator of a minimax GAN saturates and results in vanishing gradients. Our WGAN critic provides very clean gradients on all parts of the space.

- No longer have to worry about the fast learning of the discriminator.
- The gradient is smoother everywhere and learns better even the generator is not producing good images.

Section 1

Introduction

1. Introduction

- We focus on the ways to measure how close $\mathbb{P}_\theta$ is to $\mathbb{P}_r$;

$$\rho(\mathbb{P}_\theta, \mathbb{P}_r)$$

- The most fundamental difference: Their impact on the convergence of sequence of probability distribution.

- Note that: A sequence of distribution $(\mathbb{P}_t)_{t\in\mathbb{N}}$ **converges** $\Leftrightarrow {}^{\exists}\mathbb{P}_\infty$ s.t. $\rho(\mathbb{P}_t, \mathbb{P}_\infty)$ tends to 0.
  - We want to find a weaker metric ρ.

1. Introduction

In order to optimize the parameter θ, it is desirable to define our model distribution $\mathbb{P}_\theta$ in a manner that makes the mapping $\theta \mapsto \mathbb{P}_\theta$ continuous.

- **Continuity**: when a sequence of parameters θ_t converges to θ, the distribution $\mathbb{P}_{\theta_t}$ also converge to $\mathbb{P}_\theta$.
  - It depends on the way we compute the distance between distributions.

- The main reason we care about the mapping $\theta \mapsto \mathbb{P}_\theta$ to be continuous:
  - 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.

1. Introduction

Note that for $f : \{\theta_\alpha\} \to \{\mathbb{P}_\beta\}$, $f(\theta) = \mathbb{P}_\theta$, $f(\theta)$ is continuous if

$$^\forall \text{open } V \subset \{\mathbb{P}_\beta\},\ f^{-1}(V) \text{ is also open in } \{\theta_\alpha\}$$

For the topology on the metric space $M = (\{\mathbb{P}_\beta\}, \rho)$, $g(\theta) = \rho(\mathbb{P}_\theta, \mathbb{P}_r)$ is continuous, if $f(\theta)$ is continuous. Since for $h(\mathbb{P}_\theta) = \rho(\mathbb{P}_\theta, \mathbb{P}_r), g(\theta) = h(f(\theta))$ and distance function h is continuous. So,

$$^\forall W \in M,\ g^{-1}(V) = f^{-1}(h^{-1}(W)) \text{ is also open in } \{\theta_\alpha\}.$$

1. Introduction

논문의 기여

- GAN의 Discriminator보다 선생님 역할을 잘 할 수 있는 Critic을 사용함으로써 Gradient를 잘 전달시키고 Critic과 Generator를 최적점까지 학습할 수 있다.

- 따라서 아래와 같은 이점을 얻을 수 있다.
  - During training, you do not have to care about the balance between the discriminator and generator.
  - Mode dropping, a common problem in GAN, can be solved.

1. Introduction

Contents

Sec2. **Different Distances:** We provide a theoretical analysis of how the Earth Mover (EM) distance behaves in comparison to popular probability distances and divergences.

Sec3. **Wasserstein GAN:** we define a form of GAN called Wasserstein-GAN that minimizes a reasonable and efficient approximation of the EM distance.

Sec4. **Empirical Results**

Section 2

Different Distances

Notation

- $\mathcal{X}$: a compact metric set.
  (such as the space of images $[0,1]^d$)
- Σ : the set of all Borel subsets of $\mathcal{X}$.
- $\text{Prob}(\mathcal{X})$: the space of probability measures definded on $\mathcal{X}$.

We now define elementary distances and divergences between two distributions $\mathbb{P}_r, \mathbb{P}_g \in \text{Prob}(\mathcal{X})$.

2. Different Distances

- **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 $\mathbb{P}_r, \mathbb{P}_g$ are assumed to be absolutely continuous.

- **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 $(\mathbb{P}_r + \mathbb{P}_g)/2$

2. Different Distances

- **The Earth-Mover (EM) distance or Wasserstein-1:**

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

where $\prod(\mathbb{P}_r, \mathbb{P}_g)$ denotes the set of all joint distributions $\gamma(x, y)$ whose marginals are respectively $\mathbb{P}_r$ and $\mathbb{P}_g$.

1. Intuitively, γ (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 $\mathbb{P}_g$.
2. The EM distance then is the "cost" of the optimal transport plan.

2. Different Distances

The Earth-Mover (EM) distance or Wasserstein-1

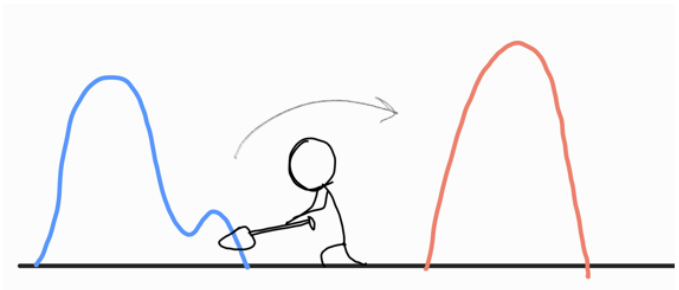
최소화

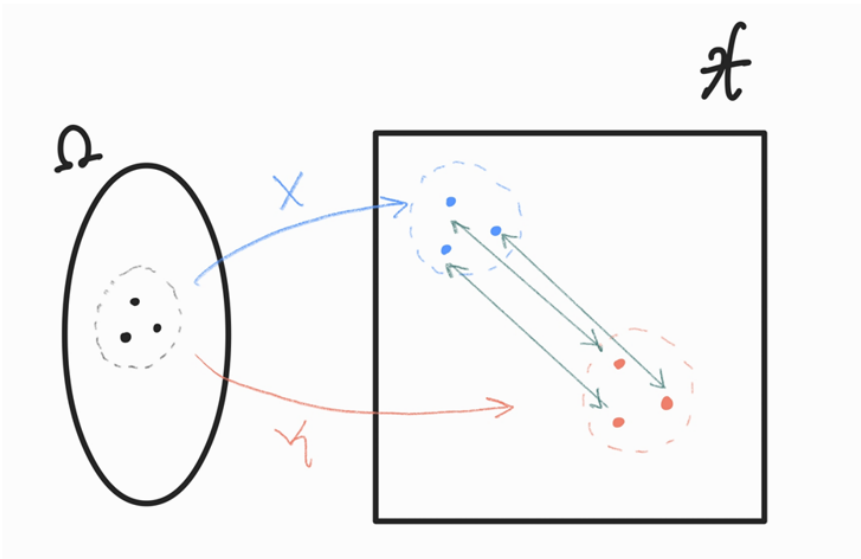
$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma \in \prod(\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(x,y)\sim\gamma}\Big[||x - y||\Big]$$

모든 전략 중 최소한의 힘

흙더미를 옮기는 모든 전략

흙더미를 옮길 때 필요한 힘





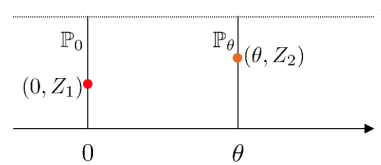
2. Different Distances

Example 1

Example 1 (Learning parallel lines)

Let $Z \sim U[0,1]$. Let $\mathbb{P}_0$ be the distribution of $(0, Z) \in \mathbb{R}^2$, Now let $g_\theta(z) = (\theta, z)$ with θ a single real parameter. In this case,

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



2. Different Distances

Outline for theoretical part

- $\theta \mapsto W(\mathbb{P}_r, \mathbb{P}_\theta)$ is continuous and almost everywhere differentiable under a certain condition.
- Hence we may expect that the estimated probability measure achieved by minimizing $W(\mathbb{P}_r, \mathbb{P}_\theta)$ with respect to θ is the more "closer" one with the true measure than any other candidates.
- Kantorovich-Rubinstein duality tells us that

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

  for each θ.
- From the duality theorem, the existence of the function attaining the supremum above is guaranteed. Saying this solution as f for a given $\theta = \theta^*$, we can derive the derivative of $W(\mathbb{P}_r, \mathbb{P}_\theta)$ at θ^* as

$$\nabla_\theta W(\mathbb{P}_r, \mathbb{P}_\theta)|_{\theta=\theta^*} = -\mathbb{E}\left[\nabla_\theta f(g_\theta(Z))|_{\theta=\theta^*}\right].$$

2. Different Distances

Coupling

Definition 1.1 of (Villani, 2008)

Let $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ be two probability spaces. Coupling μ and ν means constructing two random variables X and Y on some probability space $(\Omega, \mathbb{P})$, such that $X \sim \mu$ and $Y \sim \nu$. The couple (X, Y) is called a coupling of (μ, ν). By abuse of language, the law of (X, Y) is also called a coupling of (μ, ν).

2. Different Distances

Optimal coupling or Optimal transport

- Here one introduces a **cost function** $c(x, y)$ on $\mathcal{X} \times \mathcal{Y}$, that can be interpreted as the work needed to move one unit of mass from location x to location y.
- Then one considers the **Monge-Kantorovich minimization problem**

$$\inf \mathbb{E}c(X, Y),$$

  where the pair (X, Y) runs over all possible couplings of (μ, ν)
- Or equivalently, in terms of measures,

$$\inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) d\gamma(x, y).$$

- Such joint measures are called **transportation plans**; those achieving the infimum are called **optimal transportation plans**.

Gluing lemma (Villani, 2008)

Let $(\mathcal{X}_i, \mu_i), i = 1, 2, 3$ be Polish probability spaces. If (X_1, X_2) is a coupling of (μ_1, μ_2) and (Y_2, Y_3) is a coupling of (μ_2, μ_3), then one can construct a triple of random variables (Z_1, Z_2, Z_3) such that (Z_1, Z_2) has the same law as (X_1, X_2) and (Z_2, Z_3) has the same law as (Y_2, Y_3).

Theorem 4.1 of (Villani, 2008)

Let $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ be two Polish probability spaces; let $a : \mathcal{X} \to \mathbb{R} \cup \{-\infty\}$ and $b : \mathcal{Y} \to \mathbb{R} \cup \{-\infty\}$ be two upper semicontinuous functions such that $a \in L^1(\mu), b \in L^1(\nu)$. Let $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ be a lower semicontinuous cost function, such that $c(x, y) \geq a(x) + b(y)$ for all x, y. Then there is a coupling of (μ, ν) which minimizes the total cost $\mathbb{E}c(X, Y)$ among all possible couplings (X, Y).

2. Different Distances

Wasserstein distance is indeed a distance

- Obviously symmetric.
- Existence theorem for optimal coupling implies the identity of indiscernibles.
- Gluing lemma can be applied to show the triangle inequality.

2. Different Distances

Continuity

Theorem 1

Let $\mathbb{P}_r$ be a fixed distribution over a compact metric space $\mathcal{X}$. Let Z be a random variable over another space $\mathcal{Z}$. Let $g : \mathcal{Z} \times \mathbb{R}^d \to \mathcal{X}$ be a function, that will be denoted as $g_\theta(z)$. Let $\mathbb{P}_\theta$ denote the distribution of $g_\theta(Z)$. Then

1. If g is continuous in θ, so is $W(\mathbb{P}_r, \mathbb{P}_\theta)$.
2. If g is locally Lipschitz with local Lipschitz constants $L(\theta, z)$ satisfying $\mathbb{E}[L(\theta, Z)] < +\infty$, then $W(\mathbb{P}_r, \mathbb{P}_\theta)$ is continuous everywhere, and differentiable almost everywhere.
3. Statements 1 and 2 are false for the Jensen-Shannon divergence and all the KLs.

Corollary 1

Let g_θ be any feedforward neural network parametrized by θ and $\mathbb{E}[\|Z\|] < \infty$. Then conditions for Theorem 1 are satisfied.

2. Different Distances

Comparing distances and divergences of measures

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 $n \to \infty$,

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.

Section 3

Wasserstein GAN

3. Wasserstein GAN

Dual Kantorovich problem

- Recall the Monge-Kantorovich minimization problem

$$\inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y) d\gamma(x,y).$$

- Obviously this is no less than

$$\sup \left\{ \int_{\mathcal{Y}} \phi(y) d\nu(y) - \int_{\mathcal{X}} \psi(x) d\mu(x) : \phi(y) - \psi(x) \leq c(x,y) \right\}.$$

- A pair of price (ψ, ϕ) is described as **tight** if

$$\phi(y) = \inf_x \{\psi(x) + c(x,y)\}, \quad \psi(x) = \sup_y \{\phi(y) - c(x,y)\}.$$

- From an arbitrary price (ψ, ϕ) such that $\phi - \psi \leq c$, we can achieve tight pair by iterative improving procedure. Hence it is enough to restrict our attention to tight pairs for the above dual problem.

3. Wasserstein GAN

c-convexity

Definition 5.2 of (Villani, 2008)

Let $\mathcal{X}, \mathcal{Y}$ be sets, and $c : \mathcal{X} \times \mathcal{Y} \to (-\infty, +\infty]$. A function $\psi : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ is said to be c-convex if it is not identically $+\infty$, and there exists $\zeta : \mathcal{Y} \to \mathbb{R} \cup \{\pm\infty\}$ such that

$$\forall x \in \mathcal{X}, \quad \psi(x) = \sup_{y \in \mathcal{Y}} \{\zeta(y) - c(x, y)\} .$$

- Dual problem of Monge-Kantorovich minimization problem is equivalent to

$$\sup \left\{ \int_{\mathcal{Y}} \psi^c(y) d\nu(y) - \int_{\mathcal{X}} \psi(x) d\mu(x) : \psi \text{ is } c\text{-convex} \right\}$$

with

$$\psi^c(y) = \inf_x \{\psi(x) + c(x, y)\} .$$

3. Wasserstein GAN

Kantorovich-Rubinstein duality

Theorem 5.10 in (Villani, 2008)

Let $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ be two Polish probability spaces and let $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ be a lower semicontinuous cost function, such that

$$\forall (x, y) \in \mathcal{X} \times \mathcal{Y}, \quad c(x, y) \geq a(x) + b(y)$$

for some real-valued upper semicontinuous functions $a \in L^1(\mu)$ and $b \in L^1(\nu)$. Then the infimum of the primal problem coincides with the supremum of the dual problem.
If further c is real-valued, $C(\mu, \nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \int c d\gamma < +\infty$, and one has the pointwise upper bound

$$c(x, y) \leq c_{\mathcal{X}}(x) + c_{\mathcal{Y}}(y), \quad (c_{\mathcal{X}}, c_{\mathcal{Y}}) \in L^1(\mu) \times L^1(\nu),$$

then both the primal and dual Kantorovich problems have solutions.

3. Wasserstein GAN

Wasserstein distance

- If c is a **distance** on some metric space $\mathcal{X}$, then a c-convex function is just a 1-Lipschitz function, and it is its own c-transform.
- Thus we have

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

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 the condition in the previous theorem. Then for each $\theta = \theta^*$ there is a solution $f : \mathcal{X} \to \mathbb{R}$ to the above problem and we have

$$\nabla_\theta W(\mathbb{P}_r, \mathbb{P}_\theta)|_{\theta=\theta^*} = -\mathbb{E}\left[\nabla_\theta f(g_\theta(Z))|_{\theta=\theta^*}\right].$$

3. Wasserstein GAN

정리

$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma \in \prod(\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(x,y)\sim\gamma}\Big[||x - y||\Big] \tag{1}$$

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

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

where $\{f_w\}_{w\in\mathcal{W}}$ are all K-Lipschitz for some K.

3. Wasserstein GAN

정리

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

$$K \cdot W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma \in \prod(\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(x,y)\sim\gamma}\Big[K \cdot ||x - y||\Big]$$

Then,

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

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

by Kantorovich-Rubinstein duality.

- If we replace $||f||_L \leq 1$ for $||f||_L \leq K$ (K-Lipschitz for some constant K), then we end up with $K \cdot W(\mathbb{P}_r, \mathbb{P}_\theta)$.

3. Wasserstein GAN

정리

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

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

- This process would yield a calculation of $W(\mathbb{P}_r, \mathbb{P}_\theta)$.
- We could consider differentiating $W(\mathbb{P}_r, \mathbb{P}_\theta)$ by back-prop through (2) via estimating $\mathbb{E}_{z \sim p(z)}[\nabla_\theta f_w(g_\theta(z))]$

- f_w: K-Lipschitz --→ 'Weight Clipping'

3. Wasserstein GAN

Algorithm

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.

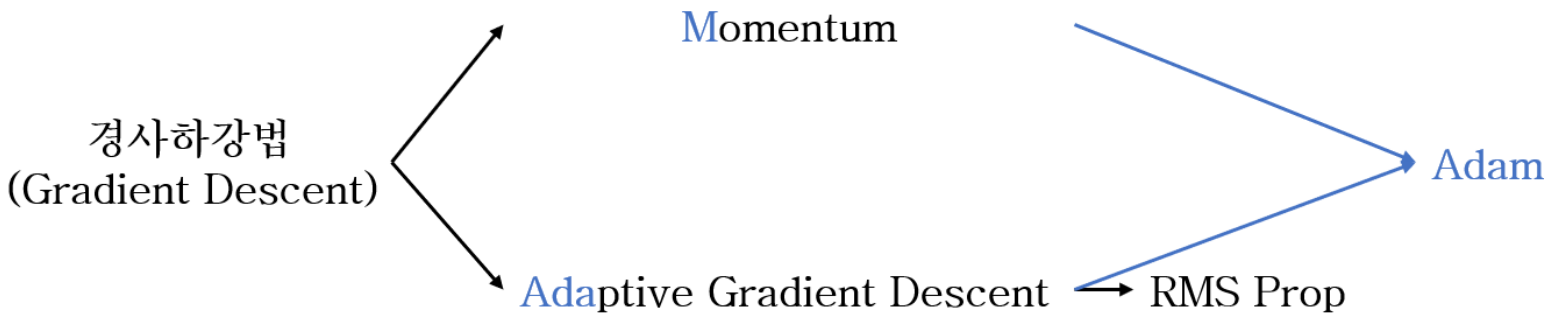
while θ has not converged **do**
  for $t = 0, ..., n_{\text{critic}}$ **do**
    Sample $\{x^{(i)}\}_{i=1}^{m} \sim \mathbb{P}_r$ a batch from the real data.
    Sample $\{z^{(i)}\}_{i=1}^{m} \sim p(z)$ a batch of prior samples.
    $g_w \leftarrow \nabla_w \left[\frac{1}{m}\sum_{i=1}^{m} f_w(x^{(i)}) - \frac{1}{m}\sum_{i=1}^{m} f_w(g_\theta(z^{(i)}))\right]$
    $w \leftarrow w + \alpha \cdot \text{RMSProp}(w, g_w)$
    $w \leftarrow \text{clip}(w, -c, c)$
  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

3. Wasserstein GAN

RMSProp

$$w \leftarrow w + \alpha \cdot \mathrm{RMSProp}(w, g_w)$$

$$\theta \leftarrow \theta - \alpha \cdot \mathrm{RMSProp}(\theta, g_\theta)$$



- Since the loss for the critic is nonstationary, momentum based methods(such as *Adam*) seemed to perform worse.
- We therefore switched to *RMSProp* which is known to perform well even on very nonstationary problems.

3. Wasserstein GAN

Clipping Issue - Lipschitz constraint

$$w \leftarrow \text{clip}(w, -c, c)$$

- $|f(x_1) - f(x_2)| \leq K|x_1 - x_2|$
- Note that the fact:

  $\mathcal{W}$ is compact $\Rightarrow$ all f_w will be K-Lipschitz for some K

- Simple we can do is clamp the weights to a fixed box, say

  $$\mathcal{W} = [-0.01, 0.01]^l,$$

  after each gradient update.

  - $\text{clamp}(x, a, b) = \begin{cases} b & \text{if } x \geq b \\ a & \text{if } x \leq a \\ x & \text{o.w.} \end{cases}$

3. Wasserstein GAN

Clipping Issue - Lipschitz constraint

$\mathcal{W}$ is compact $\Rightarrow$ all f_w will be K-Lipschitz for some K

- w is weight of neural network, composition of activation functions and linear transformations.

$$f_w(x) = \sigma(wx + \text{bias})$$

- Activation functions(sigmoid, Relu, tanh), σ, are 1-Lipschitz function.
- So Lipschitz constant K of neural network depends on value of w (If it has n multiple layer, roughly saying, it depends on w^n)

$$|f_w(x) - f_w(y)| = |\sigma(wx + b) - \sigma(wy + b)| \leq 1 \cdot |wx - wy|$$

- So if we constrain w to lie in compact space $\mathcal{W}$, closed and bounded space, Lipschitz constant K would be decided.

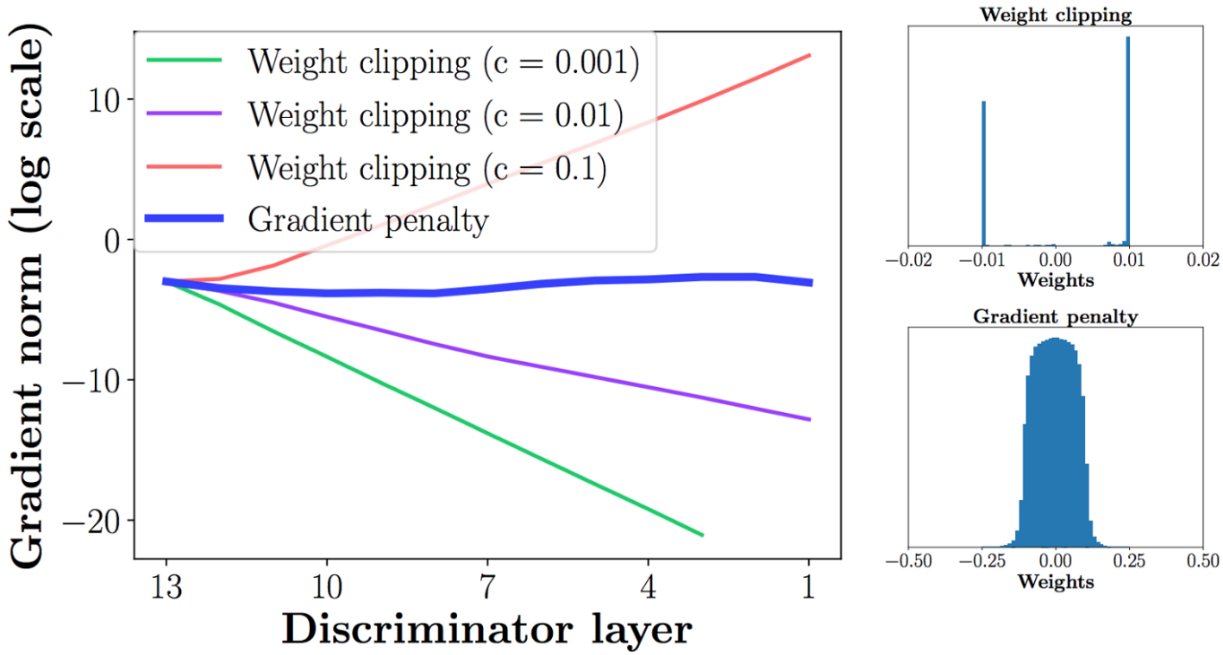
3. Wasserstein GAN

Clipping Issue - Lipschitz constraint

> Weight clipping is a clearly terrible way to enforce a Lipschitz constraint. If the clipping parameter is large, then it can take a long time for any weights to reach their limit ··· If the clipping is small, this can easily lead to vanishing gradients ··· we stuck with weight clipping due to its simplicity and already good performance. ··· However, we do leave the topic of enforcing Lipschitz constraints in a neural network setting for further investigation, and we actively encourage interested researchers to improve on this method.

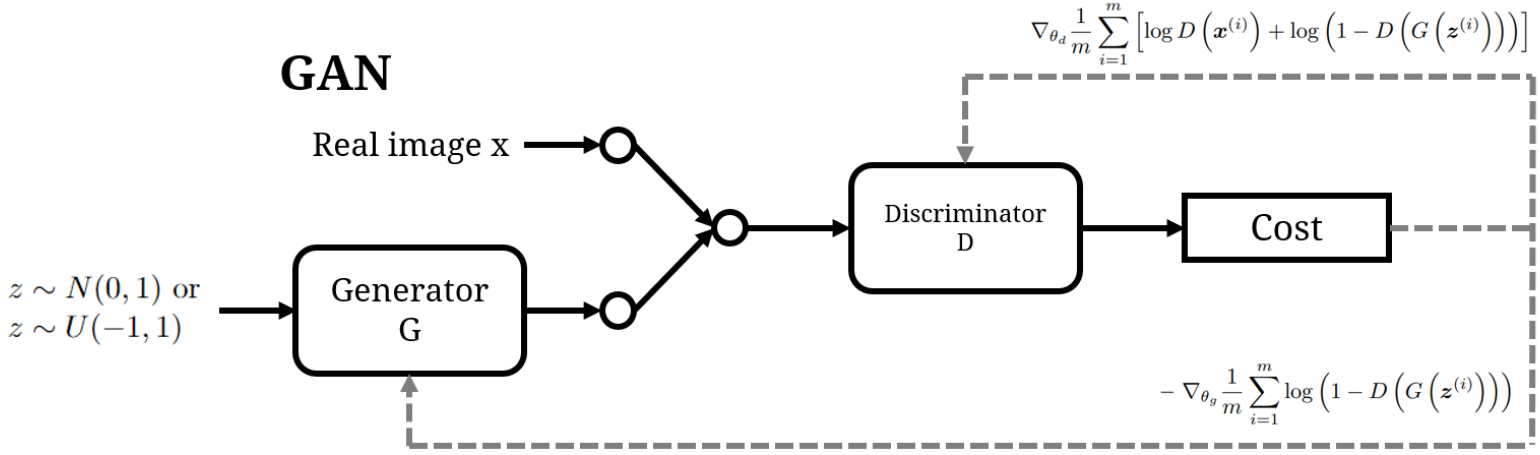
3. Wasserstein GAN

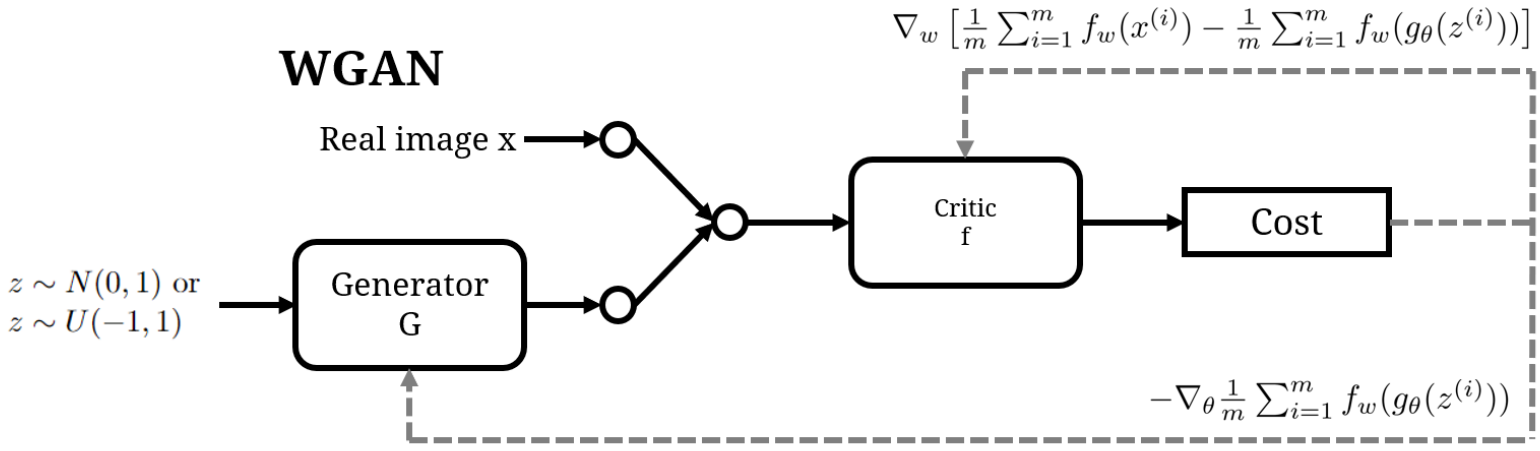
Clipping Issue - Lipschitz constraint



- The model performance is very sensitive to this hyperparameter.(batch normalization is off)
- Instead of applying clipping, **WGAN-GP** penalizes the model if the gradient norm moves away from its target norm value 1.

GAN과 WGAN





	Discriminator/Critic	Generator
GAN	$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} [\log D(x^{(i)}) + \log(1 - D(G(z^{(i)})))]$	$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log(D(G(z^{(i)})))$
WGAN	$\nabla_w \frac{1}{m} \sum_{i=1}^{m} [f(x^{(i)}) - f(G(z^{(i)}))]$	$\nabla_\theta \sum_{i=1}^{m} \sum_{i=1}^{m} f(G(z^{(i)}))$

Section 4

Empirical Results

4. Empirical Results

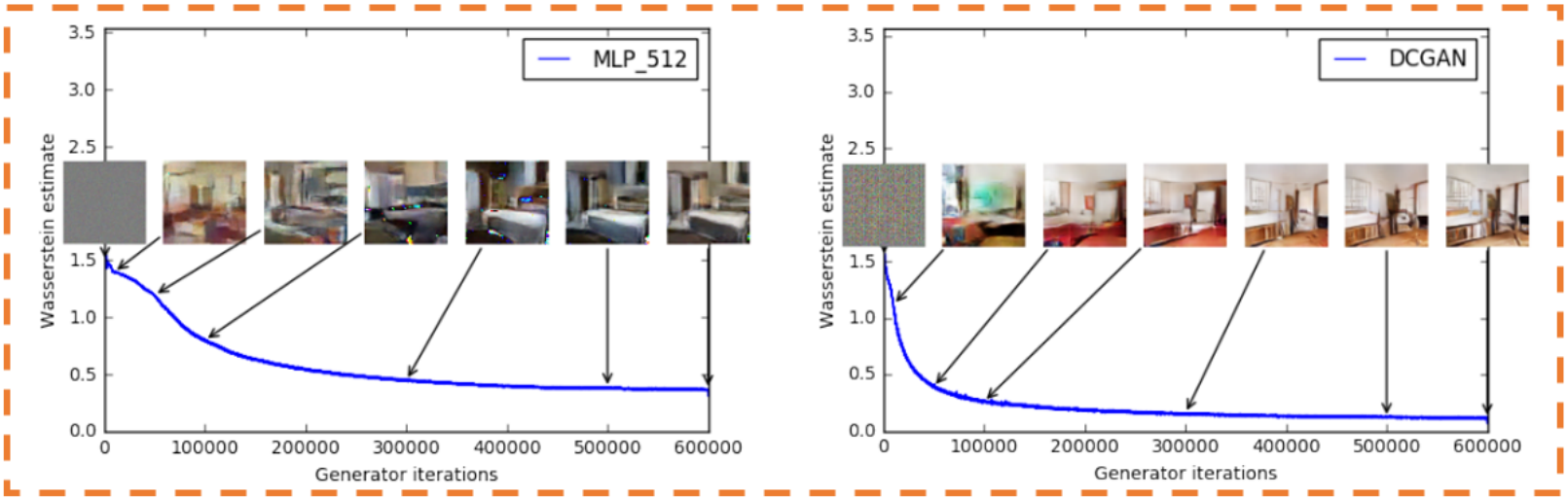
We claim two main benefits:

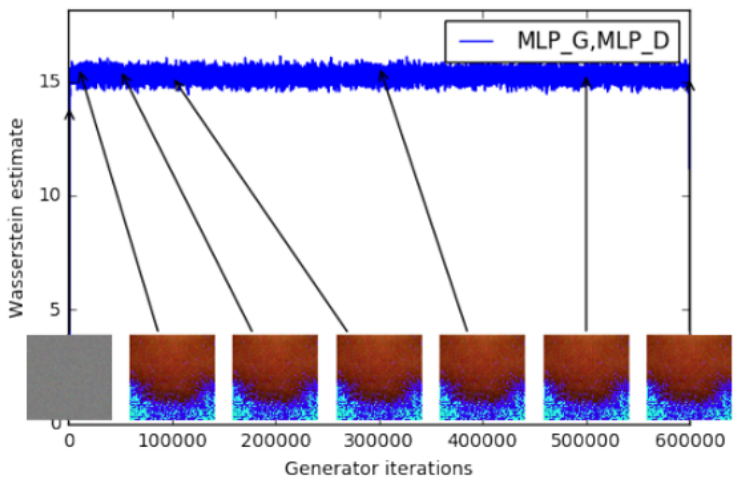
- A meaningful loss metric that correlates with the generator's convergence and sample quality
- Improved stability of the optimization process

4. Empirical Results

4.1 Experimental Procedure

discriminator대신에 critic을 적용

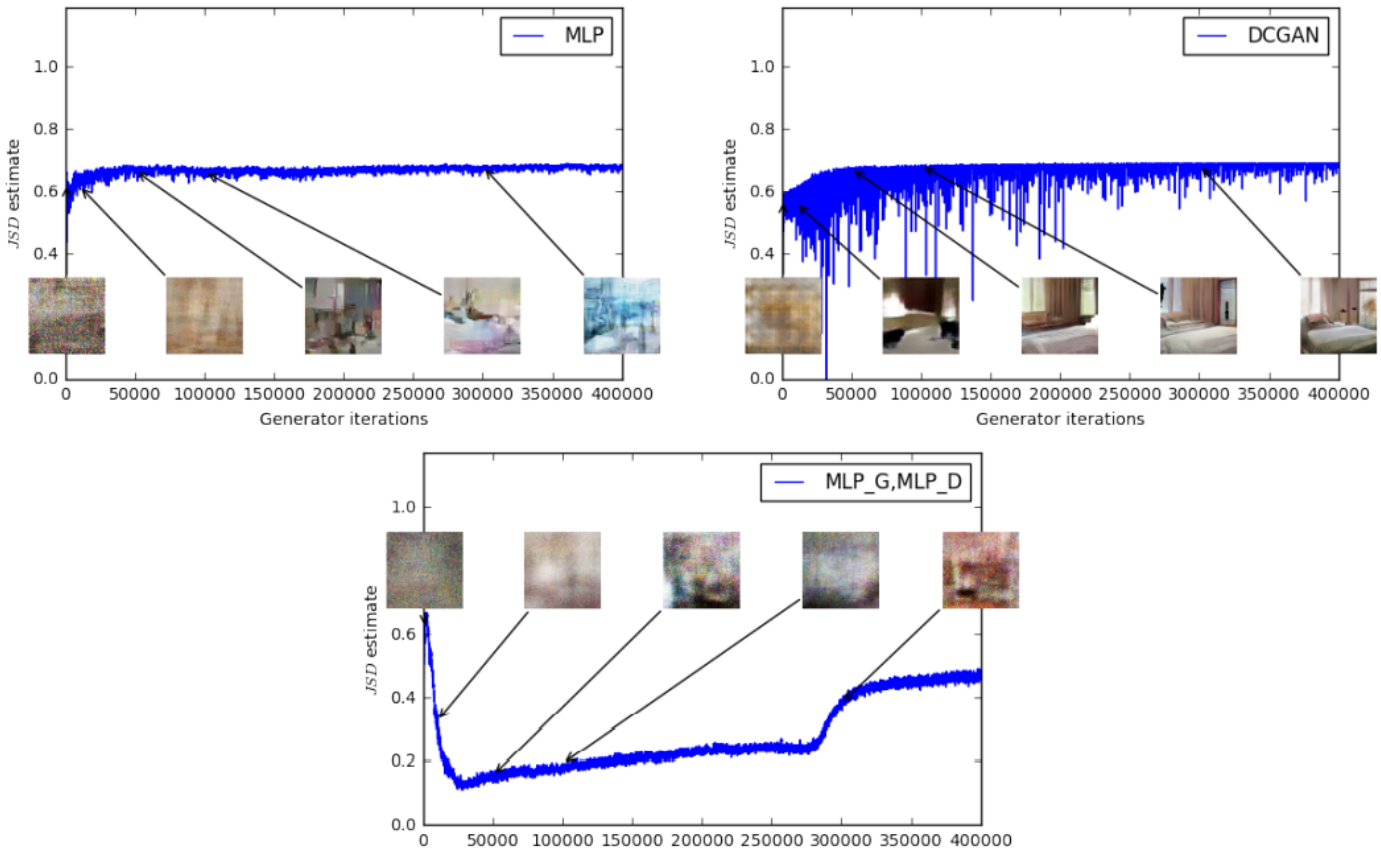




- The loss decreases quickly and sample quality increases as well.

4. Empirical Results

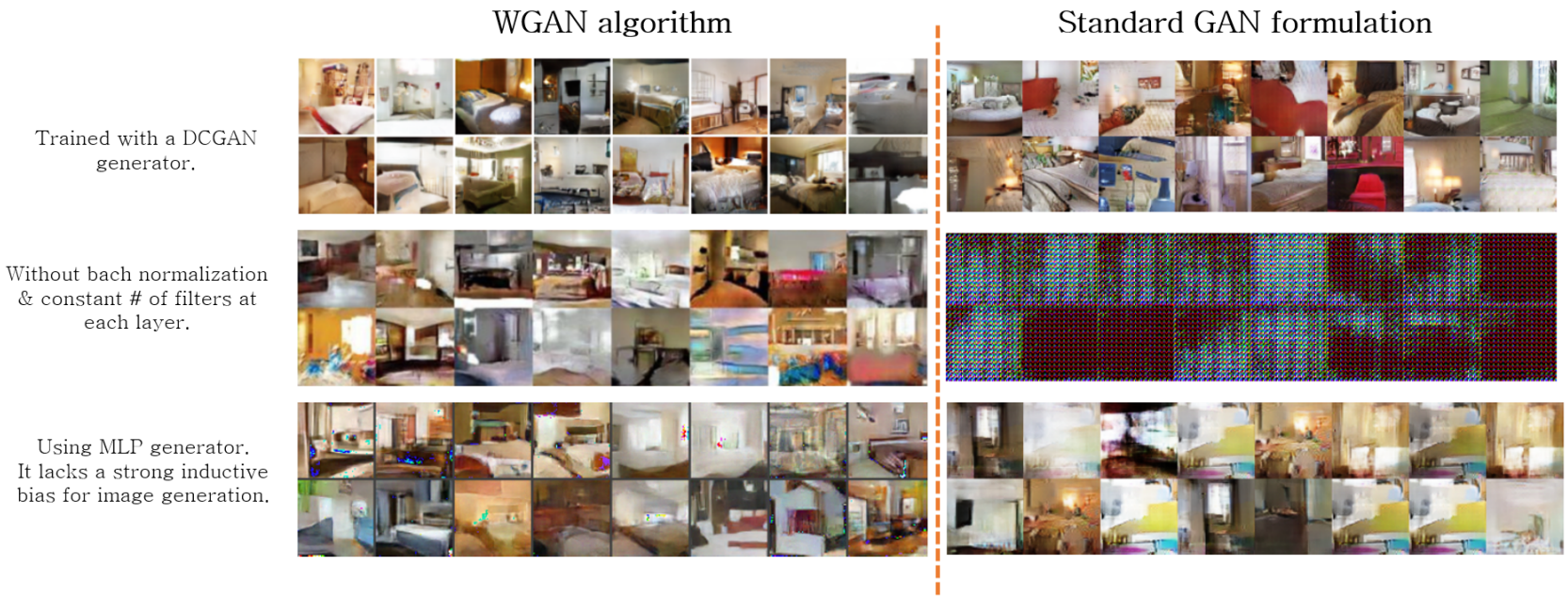
4.2 Meaningful loss metric



- JS estimates for same setting.
- We have succesfully used the loss metric to validate our experiments repeatedly

4. Empirical Results

4.3 Improved stability



- Even though we remove the batch normalization in DCGAN, WGAN can still perform.
- It has no sign of mode collapse in experiments, and the generator can still learn when the critic perform well.

감사합니다.