Understanding Generative Adversarial Networks

Hao Dong

Peking University

So far

- GAN is a couple of Generator and Discriminator; its training process is a min-max game as follows:
  - $\min_G \max_D V(D,G) = \min_G \max_D \mathbb{E}_{\boldsymbol{x} \sim p_{data}}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z} \sim p_z}[\log(1 - D(G(\boldsymbol{z}))]$
  - Theoretical guarantee: This min-max game has a global optimum for $p_g = p_{data}$
  - However there remains some fundamental problems of GAN training.
- Note that when we say “manifold P” where P is indeed a probability distribution, we actually refer to the support set of distribution P.
- This lecture: Towards a solid understanding of GAN training.

Understanding Generative Adversarial Networks

北京大學 PEKING UNIVERSITY

- Solid Understanding of GAN Training
    - Improved Technique for Generator Loss — problems: what and why
    - Fundamental Problems of Two Types of GAN — background knowledge
    - Wasserstein Distance — some solutions
    - A Temporal Solution
- Wasserstein GAN — a super solution

- Solid Understanding of GAN Training
  - Improved Technique for Generator Loss
  - Fundamental Problems of Two Types of GAN
  - Wasserstein Distance
  - A Temporal Solution
- Wasserstein GAN

- Improved Technique for Generator Loss

- **Vanilla** Generator Loss:
  - Given $\min_G \max_D V(D,G) = \min_G \max_D \mathbb{E}_{x\sim p_{data}}[\log D(x)] + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z))]$
  - If we deduce $\mathcal{L}_D$ and $\mathcal{L}_G$ directly from min-max equation, then we get:
    - $\mathcal{L}_D = -\mathbb{E}_{\boldsymbol{x}\sim p_{data}}[\log D(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{z}\sim p_z}\left[\log(1 - D\left(G(\boldsymbol{z})\right))\right]$
    - $\mathcal{L}_G = E_{z\sim p_z}[\log(1 - D(G(z)))]$ (Vanilla GAN)
  - In early training stage: Vanishing Gradient
    - D is easy to distinguish generated sample $G(z)$ from real images x

- Improved Technique for Generator Loss

- **Improved** Generator Loss:
  - If we deduce $\mathcal{L}_G$ directly from min-max equation, then we get:
    - $\mathcal{L}_G = E_{z\sim p_z}[\log(1 - D(G(z)))]$ (Vanilla GAN)
  - Known $|\nabla \log(x)| = |\frac{1}{x}|$ is significantly larger than $|\nabla \log(1-x)| = |\frac{1}{x-1}|$
  - It is the same: $\mathcal{L}_G{}' = -E_{z\sim p_z}[\log(D(G(z)))]$ (Improved GAN)
    - Minimising $\mathcal{L}_G{}'$ is equivalent to minimise $\mathcal{L}_G$, while providing larger gradient for the generator in early stage training.

$$G^* = \max_G \mathbb{E}_{\boldsymbol{z}\sim p_z}[\log D(G(\boldsymbol{z}))]$$
$$= \min_G \mathbb{E}_{\boldsymbol{z}\sim p_z}[\log(1 - D(G(\boldsymbol{z})))]$$

  - Also have

$$\min_G \max_D V(D,G) = \min_G \max_D \mathbb{E}_{x\sim p_{data}}[\log D(x)] + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z))]$$

北京大學
PEKING UNIVERSITY

- Solid Understanding of GAN Training
  - Improved Technique for Generator Loss
  - Fundamental Problems of Two Types of GAN
  - Wasserstein Distance
  - A Temporal Solution
- Wasserstein GAN

Fundamental Problems of Two Types of GAN

- In the following slides, we denote GAN with improved generator loss as Improved GAN.

- Then we claim that these two types of GAN suffer from some fundamental problems respectively:

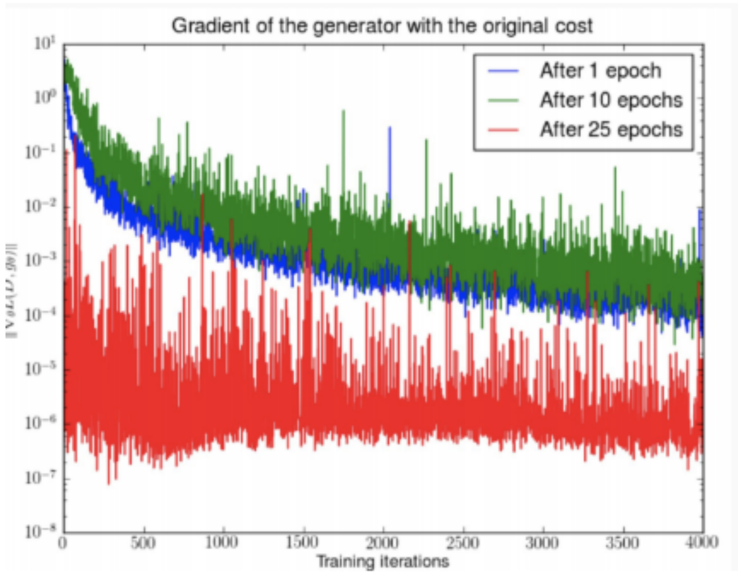
  - Vanilla GAN: *Vanishing Gradient*

  - Improved GAN: *Mode collapse and Oscillations*

Fundamental Problems of Two Types of GAN

- Vanilla GAN: Vanishing Gradient
- Improved GAN: Oscillations and Model Collapse

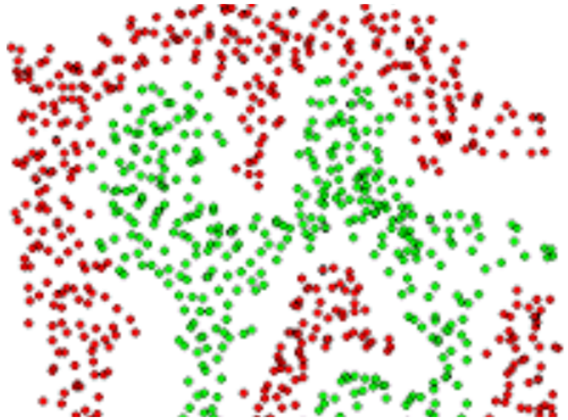
- An Empirical Observation v.s. Theoretical Induction:
  - What would happen if we just train D till converge?
    - Theoretically:
      - $D^* = \frac{p_{data}}{p_g + p_{data}}$
      - $L_G = -log4 + 2JS(p_{data}||p_g)$
    - Empirically, no gradient for G: **Why?**
      - 1. $D^*(x) = \begin{cases} 0 \ if \ x \ sampled \ from \ P_r \\ 1 \ if \ x \ sampled \ from \ P_g \end{cases}$
      - 2. $L_G = 0$
      - 3. $\nabla_x E_{z\sim p_z}[\log(1 - D^*(G(z)))] \approx 0$ (Gradient Vanishing)



Fundamental Problems of Two Types of GAN

- Vanilla GAN: Vanishing Gradient
- Improved GAN: Oscillations and Mode Collapse

- Based on empirical observations, we can intuitively thinking:
  - In what case can we classify two manifolds totally?
    - Two manifolds can be separated?
  - Consider the extreme case:
    - When support sets of P_r, P_g can be separated:
    - Then for any $x \in P_r \cup P_g$, there're only 2 cases:
      - $1.P_r(x) = 0, P_g(x) \neq 0$
      - $2.P_r(x) \neq 0, P_g(x) = 0$
    - In both case the $JS(P_r||P_g) = 2 * \frac{1}{2} * log2 = log2$
    - So $L_G = 2JS(P_r||P_g) - log4 = 0$



Fundamental Problems of Two Types of GAN

- Vanilla GAN: Vanishing Gradient
- Improved GAN: Oscillations and Mode Collapse

- Under the assumption that P_r and P_g can be separated, we can explain the reason.
  - But why?
- Firstly, it's reasonable to assume that P_r and P_g are low-dimension manifolds.
  - **Lemma 1.** *Let $g : \mathcal{Z} \to \mathcal{X}$ be a function composed by affine transformations and pointwise nonlinearities, which can either be rectifiers, leaky rectifiers, or smooth strictly increasing functions (such as the sigmoid, tanh, softplus, etc). Then, $g(\mathcal{Z})$ is contained in a countable union of manifolds of dimension at most $\dim \mathcal{Z}$. Therefore, if the dimension of $\mathcal{Z}$ is less than the one of $\mathcal{X}$, $g(\mathcal{Z})$ will be a set of measure 0 in $\mathcal{X}$.*
    - So P_g is low-dimension manifold.
  - There is strong.
- empirical and theoretical evidence to believe that P_r is indeed extremely concentrated on a low dimensional manifold

Fundamental Problems of Two Types of GAN

- Vanilla GAN: Vanishing Gradient
- Improved GAN: Oscillations and Mode Collapse

- Intuitively, when P_r and P_g are both low-dimensional, then they have "nearly no intersection" with a probability of 1.
  - The following lemma claim the same idea.
  - **Lemma 2.** *Let $\mathcal{M}$ and $\mathcal{P}$ be two regular submanifolds of $\mathbb{R}^d$ that don't have full dimension. Let η, η' be arbitrary independent continuous random variables. We therefore define the perturbed manifolds as $\tilde{\mathcal{M}} = \mathcal{M} + \eta$ and $\tilde{\mathcal{P}} = \mathcal{P} + \eta'$. Then*

$$\mathbb{P}_{\eta,\eta'}(\tilde{\mathcal{M}} \text{ does not perfectly align with } \tilde{\mathcal{P}}) = 1$$

Fundamental Problems of Two Types of GAN

- Vanilla GAN: Vanishing Gradient
- Improved GAN: Oscillations and Mode Collapse

- Further, if the 2nd order Lipschitz factor of the generator function is bounded, then as discriminator updates closer to the optimum, the generator's gradients vanishes.
  - The following lemma claim the same idea.
  - **Theorem 2.4 (Vanishing gradients on the generator).** *Let $g_\theta : \mathcal{Z} \to \mathcal{X}$ be a differentiable function that induces a distribution $\mathbb{P}_g$. Let $\mathbb{P}_r$ be the real data distribution. Let D be a differentiable discriminator. If the conditions of Theorems 2.1 or 2.2 are satisfied, $\|D - D^*\| < \epsilon$, and $\mathbb{E}_{z\sim p(z)}\left[\|J_\theta g_\theta(z)\|_2^2\right] \le M^2$, [2] then*

$$\|\nabla_\theta \mathbb{E}_{z\sim p(z)}[\log(1 - D(g_\theta(z)))]\|_2 < M\frac{\epsilon}{1-\epsilon}$$

- So far, all below questions are answered.
  - 1. $D^*(x) = \begin{cases} 0\ if\ x\ sampled\ from\ P_r \\ 1\ if\ x\ sampled\ from\ P_g \end{cases}$
  - 2. $L_G = 0$
  - 3. $\nabla_x E_{z\sim p_z}[\log(1 - D^*(G(z)))] \approx 0$ (Gradient Vanishing)

Fundamental Problems of Two Types of GAN

- Vanilla GAN: Vanishing Gradient
- Improved GAN: Oscillations and Mode Collapse

- Just as last section, we analyze the case when D is trained to optimum:
  - 1.$L_D = E_{x\sim P_r}\left[log\left(D^*(x)\right)\right] + E_{x\sim P_g}\left[log\left(1 - D^*(x)\right)\right] = 2JS(P_r||P_g) - log4$
  - 2. $KL(P_g||P_r) = E_{P_g}\left[log\frac{\frac{P_g}{P_g+P_r}}{\frac{P_r}{P_g+P_r}}\right] = E_{x\sim P_g}\left[log\frac{1-D^*(x)}{D^*(x)}\right]$

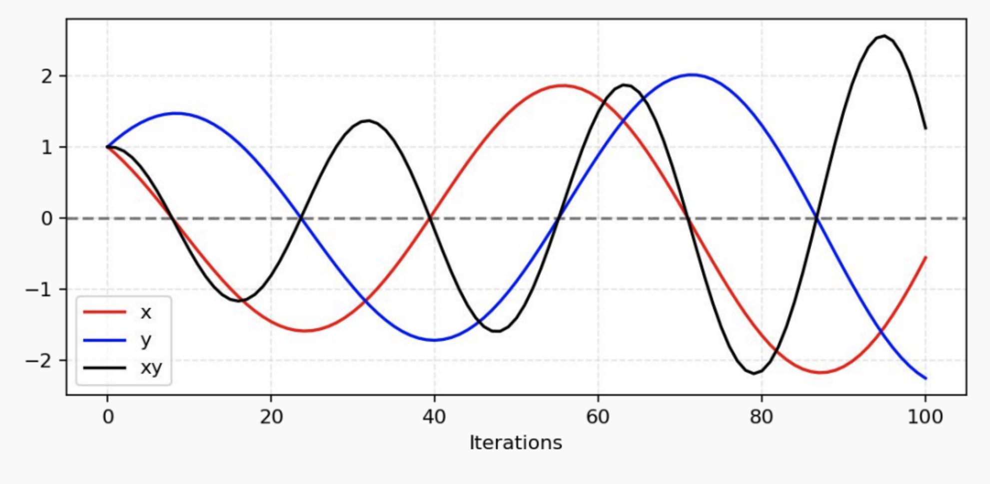
    $= E_{x\sim P_g}[1 - D^*(x)] - E_{x\sim P_g}[D^*(x)]$
- Then implied by 1. 2. :
  - $L_G = E_{x\sim P_g}[-logD^*(x)] = KL(P_g||P_r) - E_{x\sim P_g}log(1 - D^*(x))$ [implied by 2.]

    $= KL(P_g||P_r) - 2JS(P_g||P_r) + log4 + E_{x\sim P_r}logD^*(x)$ [implied by 1.]
  - $min\ L_G = min\ KL(P_g||P_r) - 2JS(P_g||P_r)$

Fundamental Problems of Two Types of GAN

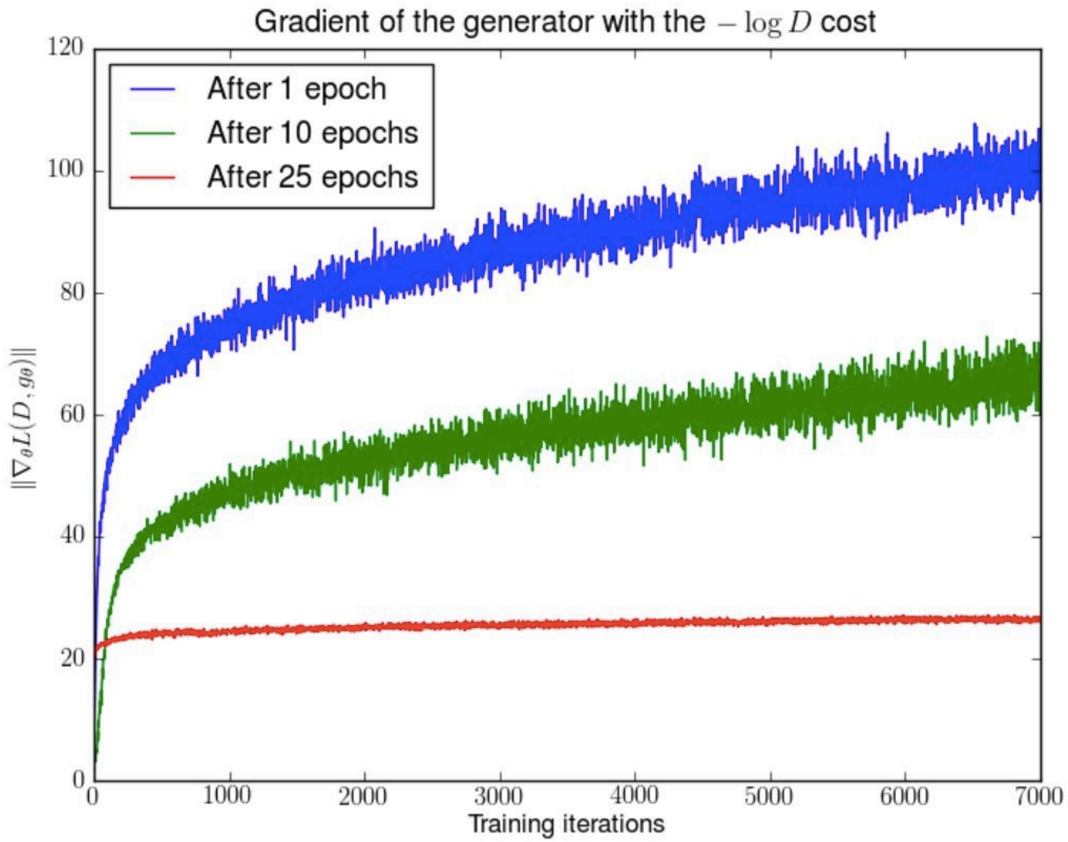
- Vanilla GAN: Vanishing Gradient
- Improved GAN: Oscillations and Mode Collapse

- $min\ L_G = min\ KL(P_g||P_r) - 2JS(P_g||P_r)$
  - Rediculous? Note that if we want to minimize L_G, then we are "pulling" $P_r\ and\ P_g$ closer and farther at the same time
  - This leads to the gradient oscillations



Fundamental Problems of Two Types of GAN

- Vanilla GAN: Vanishing Gradient
- Improved GAN: Oscillations and Mode Collapse



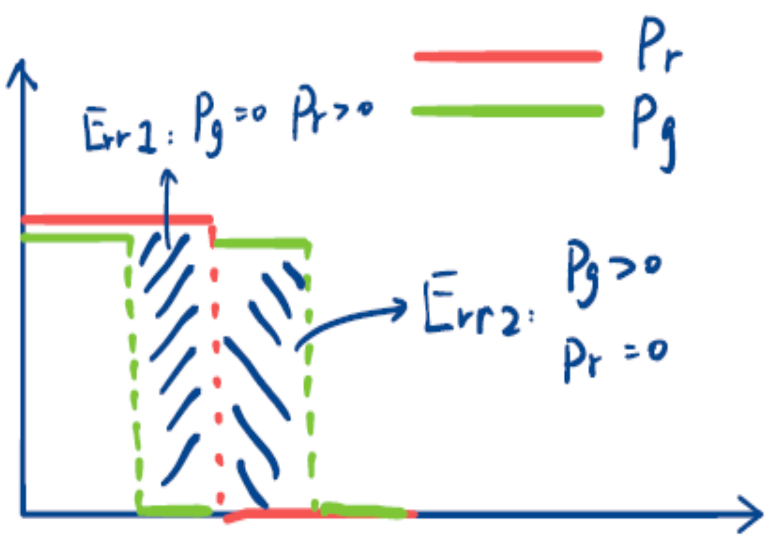
Fundamental Problems of Two Types of GAN

- Vanilla GAN: Vanishing Gradient
- Improved GAN: Oscillations and Mode Collapse

- $min\ L_G = min\ KL(P_g||P_r) - 2JS(P_g||P_r)$

- $KL(P_g||P_r) = \int P_g(x) log \frac{P_g(x)}{P_r(x)} dx$, there're two types of "error".
  - i. $P_g(x) \to 0, P_r(x) > 0$, lack of "diversity"
  - ii. $P_g(x) > 0, P_r(x) \to 0$, generate "fake" image

- Obviously, KL "punishes" type ii. more than type i.

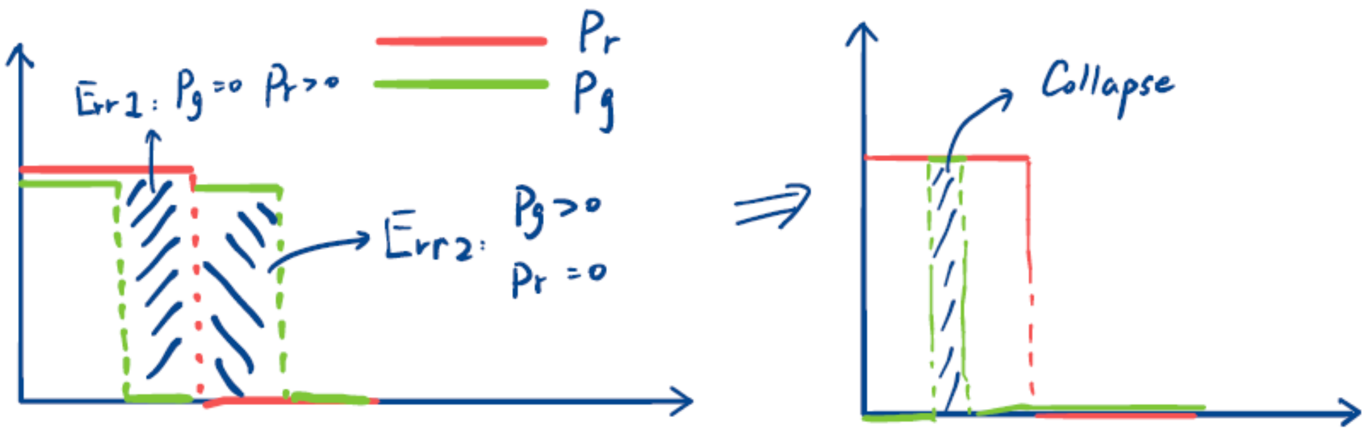


Fundamental Problems of Two Types of GAN

- Vanilla GAN: Vanishing Gradient
- Improved GAN: Oscillations and Mode Collapse

- $min\ L_G = min\ KL(P_g||P_r) - 2JS(P_g||P_r)$

- Further, to minimize $-2JS(P_g||P_r)$, error i. is "encouraged" to be more severe.



Fundamental Problems of Two Types of GAN

- Vanilla GAN: Vanishing Gradient
- Improved GAN: Oscillations and Mode Collapse

- $min\ L_G = min\ KL(P_g||P_r) - 2JS(P_g||P_r)$

- Mode collapse examples …

Fundamental Problems of Two Types of GAN

- Vanilla GAN: *Vanishing Gradient*
- Improved GAN: *Mode collapse and Oscillations*

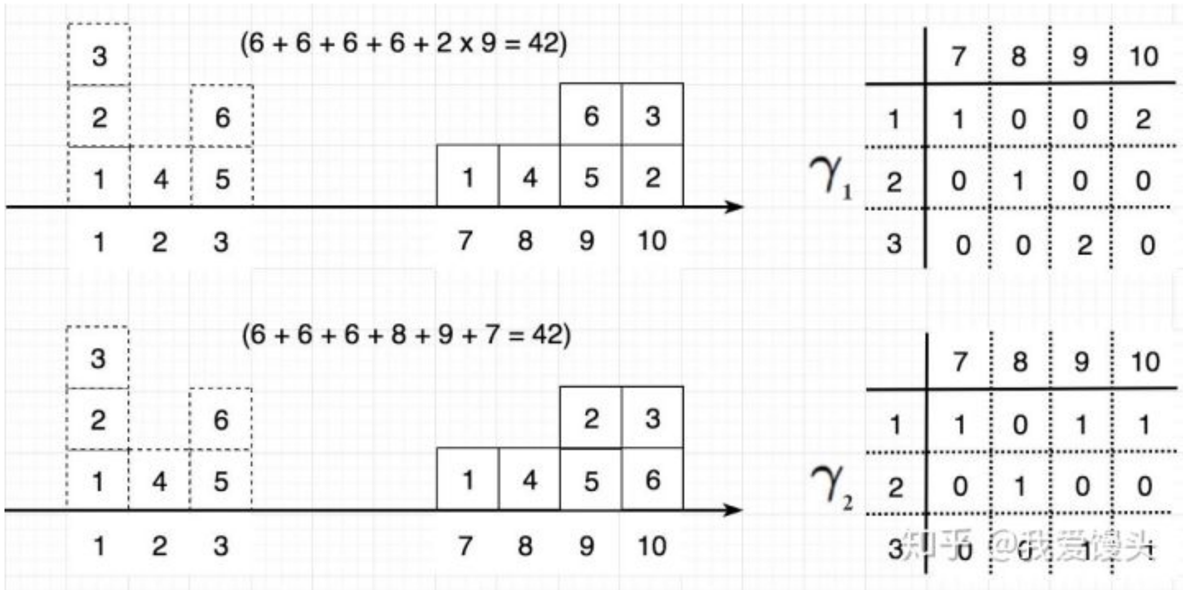
北京大學 PEKING UNIVERSITY

- Solid Understanding of GAN Training
  - Improved Technique for Generator Loss
  - Fundamental Problems of Two Types of GAN
  - Wasserstein Distance background for Wassertein GAN
  - A Temporal Solution
- Wasserstein GAN

Wasserstein Distance

- As we seen, the fundamental problem of (vanilla) GAN is due to the defects of JSD. Now we introduce a new distance.
- $W(P_r||P_g) = \inf_{\gamma \in \prod(P_g, Pr)} \mathbb{E}_{(x,y)\sim\gamma}[||x-y||]$

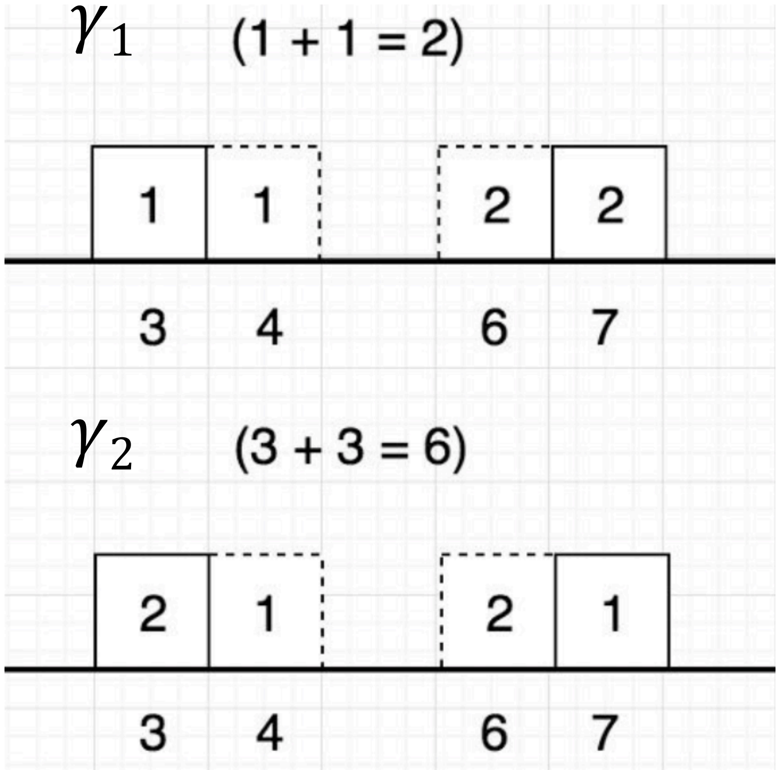
  where $\prod(Pr, Pg)$ denotes all possible joints distributions that have marginals P_r and P_g
- Wasserstein distance also goes by "earth mover's distance", the amount of "dirt" that needs to be moved to transport one distribution to the other.



Joint distribution

Joint distribution

Wasserstein Distance

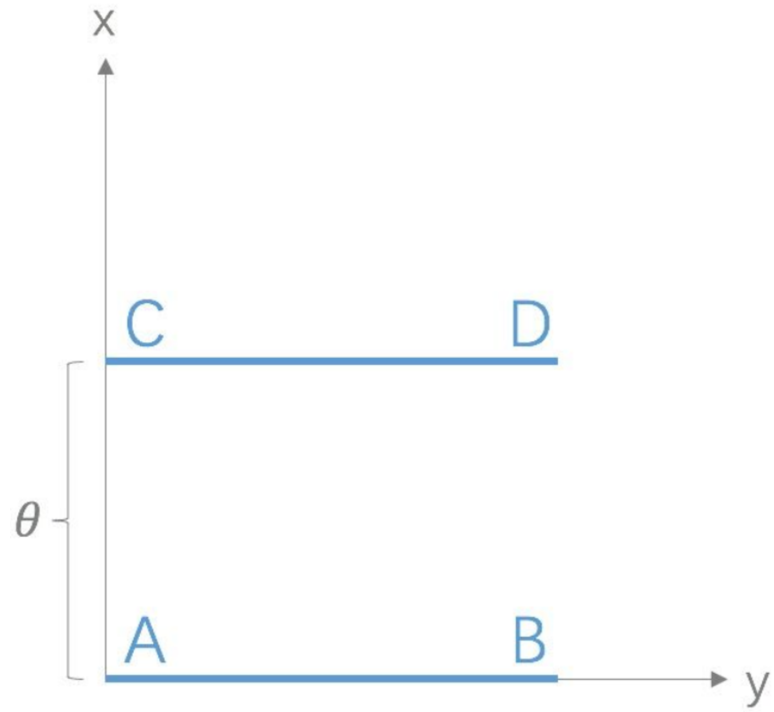


Wasserstein Distance

$$KL(P_1||P_2) = \begin{cases} +\infty & \text{if } \theta \neq 0 \\ 0 & \text{if } \theta = 0 \end{cases}$$

$$JS(P_1||P_2) = \begin{cases} \log 2 & \text{if } \theta \neq 0 \\ 0 & \text{if } \theta = 0 \end{cases}$$

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

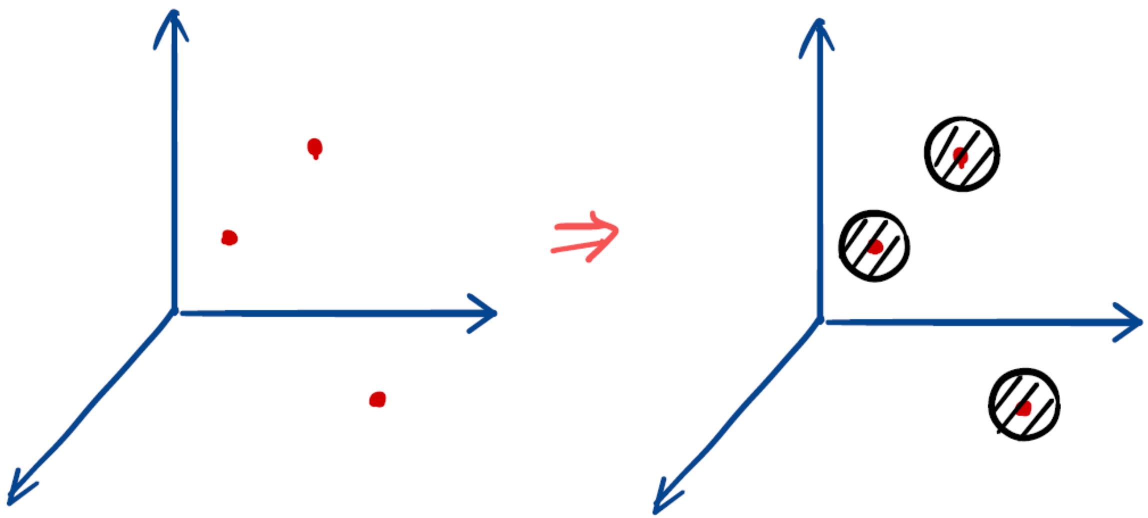


- W-distance is "better" than JSD, and JSD is "better " than KLD.
- W-distance is a better way to measure the distance between two distributions when their support sets hardly have intersection.

- Solid Understanding of GAN Training
  - Improved Technique for Generator Loss
  - Fundamental Problems of Two Types of GAN
  - Wasserstein Distance
  - A Temporal Solution
- Wasserstein GAN

A Temporal Solution: Before Wasserstein GAN

- Considering how to solve the gradient vanishing problem of Vanilla GAN
  - The problem comes from their having "nearly no intersection", due to low-dimension.
  - Idea: Add a "ϵ-ball " to each point in manifold, then a low-dimensional manifold "level-up" to full-dimensional manifold!
  - Method: Add a random vector with mean 0 and variance ϵ to each point of P_r and P_g



A Temporal Solution: Before Wasserstein GAN

PEKING UNIVERSITY 北京大學

- Relationship with Wasserstein distance
  - Let $P_{r+\epsilon}$ and $P_{g+\epsilon}$ denote the resulting manifolds respectively. Then by bounding the ϵ and $JS(P_{r+\epsilon}||P_{g+\epsilon})$, we can bound $W(P_r||P_g)$:

Theorem 3.3. *Let $\mathbb{P}_r$ and $\mathbb{P}_g$ be any two distributions, and ϵ be a random vector with mean 0 and variance V. If $\mathbb{P}_{r+\epsilon}$ and $\mathbb{P}_{g+\epsilon}$ have support contained on a ball of diameter C, then* [6]

$$W(\mathbb{P}_r, \mathbb{P}_g) \leq 2V^{\frac{1}{2}} + 2C\sqrt{JSD(\mathbb{P}_{r+\epsilon}\|\mathbb{P}_{g+\epsilon})} \tag{6}$$

- Solid Understanding of GAN Training
  - Improved Technique for Generator Loss
  - Fundamental Problems of Two Types of GAN
  - Wasserstein Distance
  - A Temporal Solution
- Wasserstein GAN

Wasserstein GAN

- **Kantorovich-Rubinstein duality**
- Lipschitz Continuity
- Wasserstein GAN

- Now we attempt to design a method to minimize the W-distance between P_r and P_g
  - $W(P_r||P_g) = \inf_{\gamma \in \prod(P_r, P_g)} \mathbb{E}_{(x,y)\sim\gamma}[||x-y||]$

Obviously, calculating the above estimation is an intractable problem.

Wasserstein GAN

- **Kantorovich-Rubinstein duality**
- Lipschitz Continuity
- Wasserstein GAN

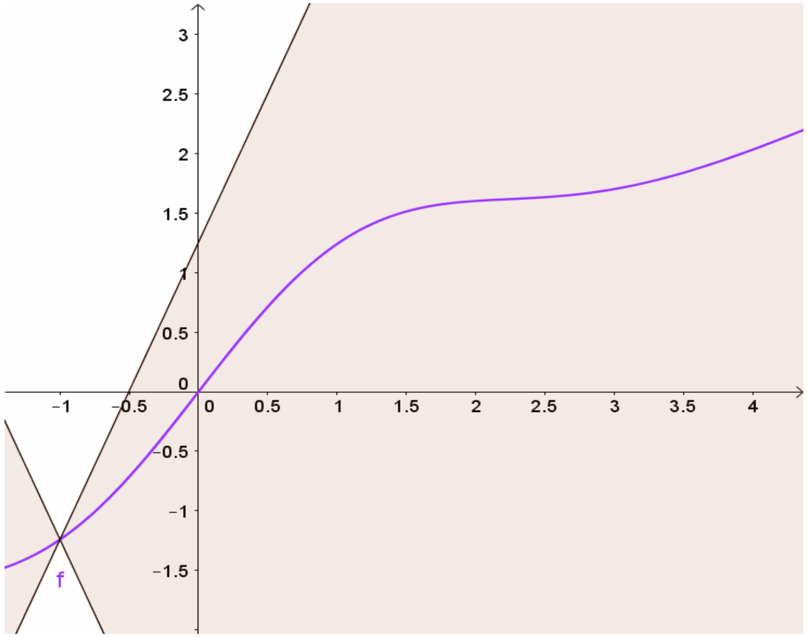
- Now we attempt to design a method to minimize the W-distance between P_r and P_g
  - Kantorovich-Rubinstein duality:
    - $W(P_r||P_g) = \frac{1}{K} \max_{||f||_L \le K} \mathbb{E}_{x \sim P_r} f(x) - \mathbb{E}_{x \sim P_g} f(x)$
    - For function f , $||f||_L$ denotes its Lipschitz-constant.

Wasserstein GAN

- Kantorovich-Rubinstein duality
- **Lipschitz Continuity**
- Wasserstein GAN

- In particular, a real-valued function $f: R^n \to R$ is called Lipschitz continuous if there exists a positive real constant K such that, for all $x_1, x_2 \in R^n$:
    - $|f(x_1) - f(x_2)| \leq K||x_1 - x_2||$

- If a function is derivable and its gradient is bounded
    - Then it is Lipschitz continuous



Wasserstein GAN

PEKING UNIVERSITY 北京大學

- Kantorovich-Rubinstein duality
- **Lipschitz Continuity**
- Wasserstein GAN

- Further, consider two functions f_1, f_2 are both Lipschitz continuous, say with constants L_1 , L_2 ,then the composition is also Lipschitz:
  - $\left|\left|f_1\left(f_2(x)\right) - f_1\left(f_2(y)\right)\right|\right| \leq L_1|f_2(x) - f_2(y)| \leq L_1 L_2||x - y||$
- So if a neural network is composed of layers that Lipschitz continuous, then the network is Lipschitz continuous.

Wasserstein GAN

- Kantorovich-Rubinstein duality
- Lipschitz Continuity
- **Wasserstein GAN**

- Now we introduce our new objective
  - To minimize $W(P_r||P_g) = \frac{1}{K} \max_{||f||_L \leq K} \mathbb{E}_{x\sim P_r} f(x) - \mathbb{E}_{x\sim P_g} f(x)$
  - Equivalent to $\min_{\mathrm{G}} W(P_r||P_g) = \frac{1}{K} \min_{\mathrm{G}} \max_{||f||_L \leq K} \mathbb{E}_{x\sim P_r} f(x) - \mathbb{E}_{x\sim P_g} f(x)$
  - Equivalent to $\min_{\mathrm{G}} W(P_r||P_g) = \min_{\mathrm{G}} \max_{||D||_L \leq K} \mathbb{E}_{x\sim P_r} D(x) - \mathbb{E}_{x\sim P_g} D(x)$

Wasserstein GAN

- Kantorovich-Rubinstein duality
- Lipschitz Continuity
- **Wasserstein GAN**

- How to optimize this objective: $\min_{G} W(P_r||P_g) = \min_{G} \max_{||D||_L \le K} \mathbb{E}_{x\sim P_r} D(x) - \mathbb{E}_{x\sim P_g} D(x)$
  - First step, fix G update D: $\max_{||D||_L \le K} \mathbb{E}_{x\sim P_r} D(x) - \mathbb{E}_{x\sim P_g} D(x)$
  - Second step, fix D update G: $\min_{G} \mathbb{E}_{x\sim P_r} D(x) - \mathbb{E}_{x\sim P_g} D(x)$
  - Obviously , the key is the first step: maximize $\mathbb{E}_{x\sim P_r} D(x) - \mathbb{E}_{x\sim P_g} D(x)$, while keeping the condition that $||D||_L \le K$

Wasserstein GAN

- Kantorovich-Rubinstein duality
- Lipschitz Continuity
- **Wasserstein GAN**

- Idea: Updating D with $\mathbb{E}_{x\sim P_r}D(x) - \mathbb{E}_{x\sim P_g}D(x)$, then clip every weight in D to $[-c, c]$ where c is a constant e.g. $c = 1$
  - After clipping, as each weight in D's each layer is bounded, then there's theorem claim that each layer is Lipschitz continuous.
  - Since each layer of D is Lipschitz continuous, then there always exists a K, such that $||f||_L \leq K$

Wasserstein GAN

- Kantorovich-Rubinstein duality
- Lipschitz Continuity
- **Wasserstein GAN**

- Algorithm:
  - 1. Sample a batch $\{x_1, x_2 \dots x_n\}, \{z_1, z_2 \dots z_n\}$
  - 2. fix G , update D with objective: $\max_D \mathbb{E}_{x\sim P_r} D(x) - \mathbb{E}_{x\sim P_g} D(x)$
  - 3.Clip every weight of D to [-1, 1]
  - 4. fix D, update G with objective: $\min_G \mathbb{E}_{x\sim P_r} D(x) - \mathbb{E}_{x\sim P_g} D(x)$
- Note that, we estimates $\mathbb{E}_{x\sim P_g} D(x) \approx \frac{1}{n}\sum_{i=1}^{n} D(G(z_i))$, $\mathbb{E}_{x\sim P_r} D(x) \approx \frac{1}{n}\sum_{i=1}^{n} D(x_i)$

Wasserstein GAN

- So … WGAN is all you need?
- In practice …
- LSGAN, WGAN-GP …

Summary: Understanding GANs

- Solid Understanding of GAN Training
  - Improved Technique for Generator Loss
  - Fundamental Problems of Two Types of GAN
  - Wasserstein Distance
  - A Temporal Solution
- Wasserstein GAN

北京大學 PEKING UNIVERSITY

Thanks