# Generative Adversarial Networks

listenzcc

April 22, 2020

**Abstract**

Not done yet.

## Contents



## 1 Real and Fake samples

Generative Adversarial Networks (GAN) can generate *FAKE* samples using adversarial learning algorithm. GAN has been widely used for extending new samples or stylize a given sample. The aim is to make the fake samples following the same distribution with *REAL* samples.

GAN has two parts, *Generator* ($G$) and *Discriminator* ($D$). The generator is to generate *FAKE* samples, the discriminator is to detect them. Thus, $G$ and $D$ are adversarial. The question is how GAN works.

## 2 Discriminator

### 2.1 Definition

A *Discriminator* is a two-classes classifier $D(x)$ that gives 1 for *REAL* sample and 0 for *FAKE* one. Thus, the loss function of a discriminator is

$$-\mathbb{E}_{x\sim P_r}[\log D(x)] - \mathbb{E}_{x\sim P_g}[\log(1 - D(x))] \tag{1}$$

where $P_r$ and $P_g$ are the probability of an image $x$ belongs to *REAL* and *FAKE* distributions.

To a given generator, the loss caused by an image is

$$\mathcal{L}(x) = -P_r(x)\log D(x) - P_g(x)\log(1 - D(x)) \tag{2}$$

## 2.2 Optimal discriminator

**Theorem 2.1.** *The optimal discriminator of a given image $x$ is*

$$D^*(x) = \frac{P_r(x)}{P_r(x) + P_g(x)} \tag{3}$$

*Proof.* Note the loss function as $\mathcal{L}$, we have

$$\begin{aligned}\frac{\partial}{\partial D}\mathcal{L} &= -\frac{P_r(x)}{D(x)} + \frac{P_g(x)}{1 - D(x)}\\ \frac{\partial}{\partial^2 D}\mathcal{L}^2 &= \frac{P_r(x)}{D^2(x)} - \frac{P_g(x)}{(1 - D(x))^2}\end{aligned}$$

One solution that minimizes the $\mathcal{L}$ is

$$D^*(x) = \frac{P_r(x)}{P_r(x) + P_g(x)}$$

when $P_r(x) \leqslant P_g(x)$, we have $\frac{\partial}{\partial^2 D}\mathcal{L}^2 \geqslant 0$. Which guarantees that $D^*(x)$ is the minimization solution.

Hence proved. □

# 3 Generator

## 3.1 Definition

The loss function of generator can be like

$$\mathbb{E}_{x\sim P_g}[\log(1 - D(x))] \tag{4}$$

or

$$\mathbb{E}_{x\sim P_g}[\log(-D(x))] \tag{5}$$

the aim is to deceive the discriminator makes $D(x) = 1$ when $x \sim P_g$.

## 3.2 Loss function under optimal discriminator

The relationship between loss function and $KL$ and $JS$ divergence is rather close.

**Theorem 3.1.** *Under optimal discriminator $D^*(x)$, the (4) equals to $JS$ divergence plus a constant.*

$$\mathbb{E}_{x\sim P_g}[\log(1 - D^*(x))] = 2JS(P_r\|P_g) - 2\log 2 - \mathbb{E}_{x\sim P_r}[\log D^*(x)]$$

*Proof.* Start by defining two measurements, $KL$ divergence and $JS$ divergence.

$$\begin{aligned}KL(P_1\|P_2) &= \mathbb{E}_{x\sim P_1}\log\frac{P_1}{P_2}\\ 2JS(P_1\|P_2) &= KL(P_1\|\frac{P_1 + P_2}{2}) + KL(P_2\|\frac{P_1 + P_2}{2})\end{aligned} \tag{6}$$

Recall (3) we have

$$\begin{aligned}\mathbb{E}_{x\sim P_r}[\log D^*(x)] &= \mathbb{E}_{x\sim P_r}[\log\frac{P_r(x)}{P_r(x) + P_g(x)}]\\ \mathbb{E}_{x\sim P_g}[\log(1 - D^*(x))] &= \mathbb{E}_{x\sim P_g}[\log\frac{P_g(x)}{P_r(x) + P_g(x)}]\end{aligned}$$

Use (6) we have

$$\mathbb{E}_{x\sim P_r}[\log D^*(x)] = KL(P_r \| \frac{P_r + P_g}{2}) - \log 2$$
$$\mathbb{E}_{x\sim P_g}[\log(1 - D^*(x))] = KL(P_g \| \frac{P_r + P_g}{2}) - \log 2$$

Adding above equations leads to

$$\mathbb{E}_{x\sim P_r}[\log D^*(x)] + \mathbb{E}_{x\sim P_g}[\log(1 - D^*(x))] = 2JS(P_r \| P_g) - 2\log 2$$

Hence proved. □

**Theorem 3.2.** *Under optimal discriminator $D^*(x)$, the (4) and (5) are related by $KL$ divergence.*

$$\mathbb{E}_{x\sim P_g}[\log(-D^*(x))] = KL(P_g \| P_r) - \mathbb{E}_{x\sim P_g}[\log(1 - D^*(x))]$$

*Proof.*

$$\begin{aligned} KL(P_g \| P_r) &= \mathbb{E}_{x\sim P_g}[\log \frac{P_g(x)}{P_r(x)}] \\ &= \mathbb{E}_{x\sim P_g}[\log \frac{1 - D^*(x)}{D^*(x)}] \\ &= \mathbb{E}_{x\sim P_g}[\log 1 - D^*(x)] - \mathbb{E}_{x\sim P_g}[\log D^*(x)] \end{aligned}$$

Hence proved. □

Use Theorem 3.1 we can conclude that under optimized discriminator, the training of the generator equals to minimize the JS divergence between $P_r$ and $P_g$.

## 3.3 Gradient vanishing

In high dimensional space, where the support set of the data *MANIFOLD* is smaller than the space. The $JS$ divergence is 0 at almost for all the images. It results that the metric is almost 0

$$\int P_r(x)P_g(x)dx \approx 0 \tag{7}$$

it shows that either $P_r$ or $P_g$ is 0 for almost every image in the space. It makes $JS$ divergence drops to 0, which means gradient vanishes. As a result, the gradient of (4) is vanishing under $D^*$.

## 3.4 Log D trick

One solution to gradient vanishing is $\log D$ trick. It changes the (4) into (5).

**Theorem 3.3.** *Minimizing (5) under optimized discriminator $D^*(x)$ is equals to minimizing following*

$$KL(P_g \| P_r) - 2JS(P_r \| P_g) \tag{8}$$

*Proof.* Combining theorem 3.1, theorem 3.2 and (5), we can proof.

Begin with theorem 3.2,

$$\mathbb{E}_{x\sim P_g}[\log(-D^*(x))] = KL(P_g\|P_r) - \mathbb{E}_{x\sim P_g}[\log(1-D^*(x))]$$

Use theorem 3.1,

$$\begin{aligned}\mathbb{E}_{x\sim P_g}[\log(-D^*(x))] = {} & KL(P_g\|P_r) - 2JS(P_r\|P_g) \\ & + 2\log 2 + \mathbb{E}_{x\sim P_r[\log(-D^*(x))]}\end{aligned}$$

It is obvious that the latter two factors is irrelevant with the generator.

Hence proved. □

It results a conflict that the optimization process requires $KL$ divergence to be smaller when $JS$ divergence to be larger at the same time. Thus, it causes unstable of the learning process.

### 3.5 Instable

The instability is mainly because of the asymmetric of the $KL$ divergence. Recall the definition of $KL$ divergence (6), we have two different situations:

ZERO: When $P_g(x) \to 0$ and $P_r(x) \to 1$, we have

$$P_g(x)\log\frac{P_g(x)}{P_r(x)} \to 0$$

INFINITY: When $P_g(x) \to 1$ and $P_r(x) \to 0$, we have

$$P_g(x)\log\frac{P_g(x)}{P_r(x)} \to +\infty$$