>>> Paper Reading: LSGAN

>>> Least Squares Generative Adversarial Networks, *ICCV 2017.*
On the Effectiveness of LSGANs, *TPAMI 2019.*

Name: 李喆琛 信息科学技术学院
初济群 数学科学学院

Date: 2020.05.14 → 第十三周·第七场

>>> Outline

1. Regular GAN and Least Squares GAN
2. Why is LSGAN better?
3. Theoretical Analysis
4. Deficiencies of LSGAN

Objective function of regular GAN

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

Why using sigmold cross entropy loss function?

Real distribution p, Fake distribution q.

- *Information entropy*: $H(p) = -\sum_i p(i) \cdot \log p(i)$;
  - $H(p) \uparrow$, Uncertainty $\uparrow$.
- *Cross entropy*: $H(p, q) = -\sum_i p(i) \cdot \log q(i)$;
  - $H(p, q) \uparrow$, Difference $\uparrow$.

The KL divergence $D(p|q) = H(p, q) - H(p)$ is a way to measure the distance between two distributions p and q.

- it comes to its minimum point when $p = q$.

Thus, GAN minimizes KL divergence.

>>> Regular Generative Adversarial Networks II

A problem: vanishing gradients

When D is closing to its optimal point D^*, the gradient of G is also clossing to zeros, that is :

$$\nabla_x E_{z\sim p(z)}[\log(1 - D(G(z)))] \approx 0$$

As P_r and P_g are two low dimension manifolds, discriminator D is easy to train, and thus closing to optimal point D^* quickly.

How to get over it ?

* Method: improved GAN
* problem: Oscillations and Mode Collapse

However, LSGAN don't have both the problems of vanilla GAN and improved GAN theorecitally.

>>> Least Squares Generative Adversarial Networks

Sigmoid cross entropy $\rightarrow$ Least squares loss function

Objective function of LSGAN

$$\min_D V_{\text{LSGAN}}(D) = \frac{1}{2}\mathbb{E}_{x\sim p_{\text{data}}(x)}\left[(D(x)-b)^2\right] + \frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}\left[(D(G(\boldsymbol{z}))-a)^2\right]$$

$$\min_G V_{\text{LSGAN}}(G) = \frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}\left[(D(G(\boldsymbol{z}))-c)^2\right]$$

The Parameters

* a: labels for fake data; Discriminator
* b: labels for real data; Discriminator
* c: values that G wants D to believe. Generator

>>> Why is LSGAN better? I

More difficult to saturate → Better stability

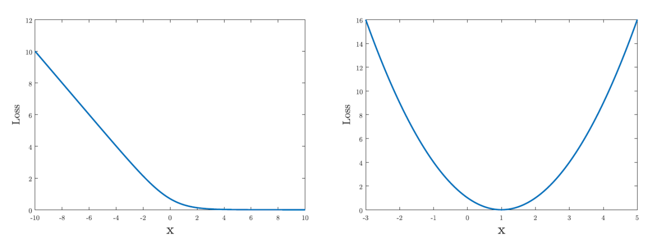


Figure: Sigmoid $y = -\log(1 - \frac{1}{1+e^{x}})$ vs. Least Square $y = (x-1)^2$

* Least squares loss function is flat only at one point;
* Sigmoid cross entropy will saturate when x is large.

>>> Why is LSGAN better? II

Tougher penalties → Higher Quailty

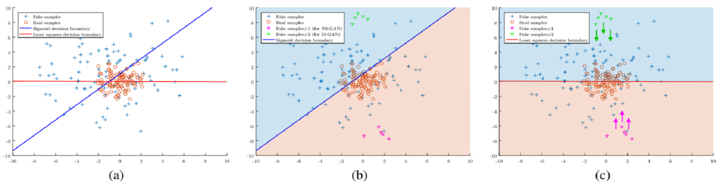
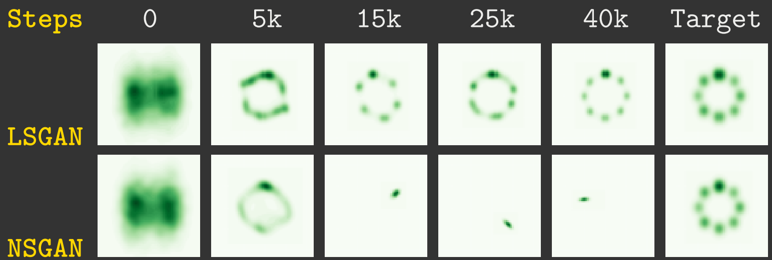


Figure: Real Samples → Orange, Fake Samples → Blue

- *Vanilla GAN*: ∇ in GREEN → Little error
  - Leads to the problem of Vanishing Gradients.
- *Improved GAN*: ☆ in PINK → Little error
  - Leads to the problem of Mode Collapse.
- *Least Squares*: Penalize samples far from the boundary.
  - Forces G to generate samples toward decision boundary.

>>> Comparison of the results



FID Results on Four Datasets

Method	LSUN	Cat	ImageNet	CIFAR10
NS-GAN	28.04	15.81	74.15	35.25
WGAN-GP	22.77	29.03	**62.05**	40.83
$\text{LSGAN}_{(011)}$	27.21	15.46	72.54	36.46
$\text{LSGAN}_{(-110)}$	**21.55**	**14.28**	68.95	**35.19**

Consider the following extensions:

$$\min_D V_{\text{LSGAN}}(D) = \frac{1}{2}\mathbb{E}_{x\sim p_{\text{data}}(x)}\left[(D(x)-b)^2\right] + \frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}\left[(D(G(\boldsymbol{z}))-a)^2\right]$$

$$\min_G V_{\text{LSGAN}}(G) = \frac{1}{2}\mathbb{E}_{x\sim p_{\text{data}}(x)}\left[(D(x)-c)^2\right] + \frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}\left[(D(G(\boldsymbol{z}))-c)^2\right]$$

* *Note that* $\mathbb{E}_{x\sim p_{\text{data}}(x)}\left[(D(x)-c)^2\right]$ *does not contain* G.

Optimal Discriminator

For a fixed G, the optimal discriminator D is:

$$D^*(x) = \frac{bp_{\text{data}}(x) + ap_g(x)}{p_{\text{data}}(x) + p_g(x)}$$

>>> Relation to Pearson χ^2 Divergence II

Proof of the optimal discriminator

In fact, we are trying to <u>minimize $V(D)$</u>:

$$\begin{aligned} V(D) &= \frac{1}{2}\mathbb{E}_{x\sim p_{\text{data}}}\left[(D(x)-b)^2\right] + \frac{1}{2}\mathbb{E}_{z\sim p_z}\left[(D(G(z))-a)^2\right] \\ &= \frac{1}{2}\mathbb{E}_{x\sim p_{\text{data}}}\left[(D(x)-b)^2\right] + \frac{1}{2}\mathbb{E}_{x\sim p_g}\left[(D(x)-a)^2\right] \\ &= \int_{\mathcal{X}} \frac{1}{2}\left(p_{\text{data}}(x)(D(x)-b)^2 + p_g(x)(D(x)-a)^2\right)\mathrm{d}x \leftarrow \text{Denoted by } Y \end{aligned}$$

Let its derivative be zero:

$$\frac{\mathrm{d}Y}{\mathrm{d}D(x)} = p_{\text{data}}(x)(D(x)-b) + p_g(x)(D(x)-a) = 0$$

Then we have:

$$D(x) = \frac{bp_{\text{data}}(x) + ap_g(x)}{p_{\text{data}}(x) + p_g(x)} \leftarrow \text{Denoted by } D^*(x)$$

In other word, $D^*(x)$ minimizes $V(D)$. □

>>> Relation to Pearson χ^2 Divergence III

Theorem. Optimizing LSGANs yields minimizing Pearson χ^2 divergence between $p_{\mathrm{d}} + p_g$ and p_g, if $b - c = 1$, and $b - a = 2$.

Proof. Substitute $D^*(x)$ into the equation:

$$
\begin{aligned}
2C(G) &= \mathbb{E}_{x\sim p_{\mathrm{d}}}\left[\left(D^*(x) - c\right)^2\right] + \mathbb{E}_{\boldsymbol{z}\sim p_z}\left[\left(D^*(G(\boldsymbol{z})) - c\right)^2\right] \\
&= \mathbb{E}_{x\sim p_{\mathrm{d}}}\left[\left(D^*(x) - c\right)^2\right] + \mathbb{E}_{x\sim p_g}\left[\left(D^*(x) - c\right)^2\right] \\
&= \mathbb{E}_{x\sim p_{\mathrm{d}}}\left[\left(\frac{bp_{\mathrm{d}}(x) + ap_g(x)}{p_{\mathrm{d}}(x) + p_g(x)} - c\right)^2\right] + \mathbb{E}_{x\sim p_g}\left[\left(\cdots\right)^2\right] \\
&= \int_{\mathcal{X}} p_{\mathrm{d}}(x)\left(\frac{(b-c)p_{\mathrm{d}}(x) + (a-c)p_g(x)}{p_{\mathrm{d}}(x) + p_g(x)}\right)^2 \mathrm{d}x + \int_{\mathcal{X}} p_g(x)\left(\cdots\right)^2 \mathrm{d}x \\
&= \int_{\mathcal{X}} \frac{\left((b-c)p_{\mathrm{d}}(x) + (a-c)p_g(x)\right)^2}{p_{\mathrm{d}}(x) + p_g(x)} \mathrm{d}x
\end{aligned}
$$

>>> Relation to Pearson χ^2 Divergence IV

Let $b - c = 1$, $b - a = 2$, we have:

$$\begin{aligned} 2C(G) &= \int_{\mathcal{X}} \frac{\left((b-c)\left(p_{\mathrm{d}}(x)+p_g(x)\right)-(b-a)p_g(x)\right)^2}{p_{\mathrm{d}}(x)+p_g(x)} \mathrm{d}x \\ &= \int_{\mathcal{X}} \frac{\left(2p_g(x)-\left(p_{\mathrm{d}}(x)+p_g(x)\right)\right)^2}{p_{\mathrm{d}}(x)+p_g(x)} \mathrm{d}x \\ &= \chi^2_{\texttt{Pearson}}\left(p_{\mathrm{d}}+p_g \| 2p_g\right) \end{aligned}$$

That proves the theorem. □

>>> Parameters Selection

Let $b-c=1,\ b-a=2 \Rightarrow$ Minimizing Pearson χ^2 Divergence

For example, $a=-1,\ b=1,\ c=0$:

$$\min_D V_{\text{LSGAN}}(D) = \frac{1}{2}\mathbb{E}_{x\sim p_{\text{data}}(x)}\left[(D(x)-1)^2\right] + \frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}\left[(D(G(\boldsymbol{z}))+1)^2\right]$$

$$\min_G V_{\text{LSGAN}}(G) = \frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}\left[(D(G(\boldsymbol{z})))^2\right]$$

Let $b=c \Rightarrow$ Generating samples as real as possible

For example, $a=0,\ b=-1,\ c=-1$:

$$\min_D V_{\text{LSGAN}}(D) = \frac{1}{2}\mathbb{E}_{x\sim p_{\text{data}}(x)}\left[(D(x)-1)^2\right] + \frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}\left[(D(G(\boldsymbol{z})))^2\right]$$

$$\min_G V_{\text{LSGAN}}(G) = \frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}\left[(D(G(\boldsymbol{z}))-1)^2\right]$$

If D is good enough, the problem still exists

Without loss of generality, let $c = 0$,
then the optimal discriminator is:

$$D^* = \frac{bP_\mathrm{d} + aP_g}{P_\mathrm{d} + P_g}$$

Plug it into the extended loss function:

$$V_\mathrm{LSGAN}(G) = \frac{1}{2}\mathbb{E}_{x\sim p_\mathrm{d}(x)}\left[D^*(x)^2\right] + \frac{1}{2}\mathbb{E}_{x\sim p_g(x)}\left[D^*(x)^2\right]$$

If **supp** p_d and **supp** p_g are *low* dimensional manifolds in *high* dimensional space, we have

$$\mathbb{P}[\mu(\textbf{supp}\ \ p_\mathrm{d} \cap \textbf{supp}\ \ p_g) = 0] = 1 \tag{1}$$

>>> Better than regular GAN, but not good enough II

For a given x, there are only four cases:

1. $p_{\mathrm{d}}(x) = 0$, $p_g(x) = 0$: We can ignore this case;
2. $p_{\mathrm{d}}(x) \neq 0$, $p_g(x) = 0$: $D^*(x) = a$, V_{LSGAN} is a constant;
3. $p_{\mathrm{d}}(x) = 0$, $p_g(x) \neq 0$: $D^*(x) = b$, V_{LSGAN} is a constant;
4. $p_{\mathrm{d}}(x) \neq 0$, $p_g(x) \neq 0$: Will not happen due to Equ (1).

Gradients vanish again!

For LSGANs, if:

* *Equ (1) holds;* (Support sets are disjoint)
* *Discriminator D is good enough;* (very close to D^*)

then the loss will be zero $\rightarrow$ Gradient vanish.

Another point of view: WGAN (Wasserstein metric)

Not Lipschitz continuous $\rightarrow$ Vanishing gradients.

To conclude, LSGAN cannot solve the problem completely.

>>> To summarize

1. *What is LSGAN*:
   1.1 Sigmoid cross entropy $\rightarrow$ Least squares;
2. *Benifits of LSGAN*:
   2.1 Better stability;
   2.2 Higher quailty;
   2.3 Partially solves the problem of vanishing gradients;
3. *Theoretical Properties*:
   3.1 Convergence: LSGAN minimizes Pearson χ^2 Divergence;
4. *Deficiencies of LSGAN*:
   4.1 Cannot solve the problem of completely.

Thanks For Listening!