Intuitive Approach to Understand the Mathematics Behind GAN

Research · March 2019

DOI: 10.13140/RG.2.2.12650.36805

S. M. Nadim Uddin
Gachon University

GAN - Intuitive Approach to Mathematics

S. M. Nadim Uddin

CVIP Lab, Gachon University

March 2019

Acknowledgement

This explanation is heavily inspired and based on [2], [6] and [3]

1 Objective function

The original paper describes the architecture with the objective function [2]:

$$\arg\min_{G}\max_{D} V(D,G) = E_{x\ p_{data}(x)}[\log(D(x))] + E_{z\ p_z(z)}[\log(1 - D(G(z)))] \quad (1)$$

where,

- $D(x)$ is the discriminator function. $D(x)$ outputs the probability that the input vector x is from the original dataset i.e. if input x is given, $D(x)$ will output a scalar value between 0 and 1.
- $G(z)$ is the generator function. $G(z)$ outputs a matrix with dimension equal to input vector x based on the noise vector z, where z can be obtained from a probability distribution.
- $P_{data}(x)$ is the probability distribution of samples of the original dataset.
- $P_z(z)$ is the probability distribution of the samples of the noise generator.
- $E(.)$ denotes *expectation function* which comes from the positive class of log-loss function. Log-loss function is defined as:

$$E(p|y) = \frac{-1}{N}\sum_{i=1}^{N}(y_i(\log p_i) + (1 - y_i)(1 - p_i)) \quad (2)$$

  where p_i is the estimation and y_i is the actual data. Log-loss function is used when expected response of the model is between 0 and 1. In other sense, $E(f(x)$ of some function $f(x)$ with respect to a probability distribution $p(x)$ is the average value of $f(x)$ when x is drawn from $p(x)$ and can be denoted as,

$$E_{x\sim p}(f(x)) = \int p(x)f(x)dx \tag{3}$$

- Equation (1) is the objective function which contains two loops denoting $\max_D V(D,G)$ and $\min_G V(D,G)$.
  1. The objective of $\max_D V(D,G)$ is to maximize the right hand side of the function $(E_{x\ p_{data}(x)}[\log(D(x))] + E_{z\ p_z(z)}[\log(1 - D(G(z)))])$ by tuning D's parameters only. Explanation is given in section 2.
  2. The objective of $\min_G V(D,G)$ is to minimize $(E_{z\ p_z(z)}[\log(1-D(G(z)))])$ by tuning G's parameters only. Note that, there is no parameter of D in the second loop and hence, $E_{x\ p_{data}(x)}[\log(D(x))]$ can be ignored. Explanation is given in section 2.

2 Detailed explanation

2.1 Optimization Problem

Recall from $\log(x)$ plot, if x becomes close to 1, $\log(x)$ becomes close to 0 and hence, $E(log(x))$ becomes close to 0. Again, when x becomes close to 0, $\log(x)$ becomes close to $-\infty$ and hence, $E(log(x))$ becomes close to $-\infty$.

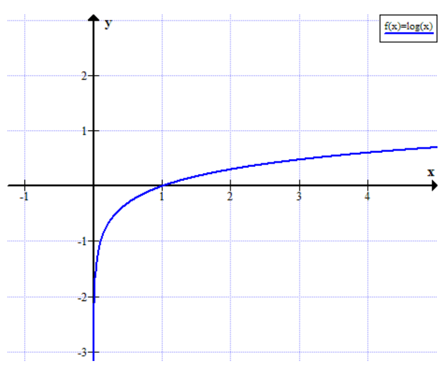

Figure 1: log(x)

Maximizing the first term of the function, i.e. $E_{x\ p_{data}(x)}[\log(D(x))]$, means $D(x)$ will try to output values close to 1 for original data. In the second term,$E_{z\ p_z(z)}[\log(1 - D(G(z)))]$, the maximum value of $\log(1 - D(G(z)$ is $+\infty$ when $D(G(z)) = 0$. So, to maximize $E_{z\ p_z(z)}[\log(1 - D(G(z)))$, $D(G(z))$ will try to output values close to 0.
Recall from equation (1), the objective of the second loop, $\min_G V(D,G)$, is to minimize $E_{z\ p_z(z)}[\log(1 - D(G(z)))]$ as the first term, $E_{x\ p_{data}(x)}[\log(D(x))]$, is

not dependent on G. The lowest value $E_{z\ p_z(z)}[\log(1 - D(G(z)))]$ can have is $-\infty$ when $D(G(z)) = 1$. Recall that D will output ~ 1 when the data is from the original dataset. It implies that $G(z)$ has to generate outputs as close as possible to the original dataset.
The optimization problem is to find a solution that will maximize $D(x)$ while minimizing $G(z)$. Maximizing $D(x)$ means that $D(x)$ will be able to properly identify real and fake(generated) data. The optimal discriminator for equation (1) with respect to a given G will be denoted as D_G^*. D_G^* can be written as,

$$D_G^* = \arg\max_D V(D, G) \tag{4}$$

However, the objective of G will be to minimize equation (1) when $D = D_G^*$. The optimal solution, denoted by G^*, can be said to satisfy

$$G^* = \arg\min_G V(D, G) \tag{5}$$

2.2 Proof of existence of optimal solution

Recall the Radon-Nikodym Theorem of measure theory [5]which states that, if there exists two σ-finite signed measures μ and ν and $\mu << \nu$, then there is a function f so that,

$$\nu(E) = \int_E f d\mu$$

By definition,

$$\nu(E) = \int_E d\nu$$

If both of the integrals are taken with respect to the same measure, then it can be written that,

$$\int_E f d\mu = \int_E g d\mu \tag{6}$$

From the equation (2) and (6), equation (1) can be written as,

$$\arg\min_G \max_D V(D, G) = \int_x p_{data}(x) \log D(x) dx + \int_x p_G(x) \log(1 - D(x)) dx \tag{7}$$

To find the optimal discriminator, it is desired to find a maximum of the integrand of equation (7). Let $p_{data}(x)$, $p_G(x)$ and $\log D(x)$ be denoted by a, b and y respectively, then the integrand can be written as,

$$f(y) = a \log y + b \log(1 - y) \tag{8}$$

To find the maximum of y,

$$f^{'}(y) = 0$$
$$\frac{a}{y} - \frac{b}{1 - y} = 0$$
$$y = \frac{a}{a + b}$$

if $a + b \neq 0$. Again,

$$f^{''}(y) = -\frac{a}{(a+b)^2} - \frac{b}{(1-\frac{a}{a+b})^2} < 0$$

when $a, b \in (0, 1)$.
So, it can be seen that $D(x) = \frac{p_{data}(x)}{p_{data}(x)+p_G(x)}$ is the maximum of the integrand. Notice that $p_{data}(x)$ is a priori and it can not be directly used during the training. However, it enables to approximate a value of D during training provided that there exists an optimal G. If $p_{data}(x) = p_G(x)$, then,

$$D_G^* = \frac{p_{data}(x)}{p_{data}(x) + p_G(x)} = \frac{1}{2} \tag{9}$$

which denotes that D will output confusing results as it will allow data from original dataset and generator. Therefore, G is solution to the mini-max game. Let $p_{data}(x) = p_G(x)$, then equation (7) becomes,

$$\begin{aligned} V(G, D_G^*) &= \int_x p_{data}(x) \log \frac{1}{2} + \int_x p_G(x) \log(1 - \frac{1}{2}) \\ &= -\log 2 \int_x p_{data}(x)dx - \log 2 \int_x p_G(x))dx \\ &= -\log 2 \int_x (p_{data}(x) + p_G(x))dx \\ &= -2\log 2 \end{aligned}$$

$$\therefore V(G, D_G^*) = -\log 4 \tag{10}$$

This is the global minimum of $C(G)$ where $C(G) = \max_D V(G, D)$ or the training criteria.

2.3 Authentication of the global minimum

From equation (7),

$$\begin{aligned} C(G) = &\int_x (\log 2 - \log 2)p_{data}(x) + p_{data}(x) \log \frac{p_d ata}{p_G(x) + p_{data}(x)} \\ &+ (\log 2 - \log 2)p_G(x) + p_G(x) \log \frac{p_G(x)}{p_G(x) + p_{data}(x)} dx \\ &= -\log 2 \int_x p_G(x) + p_{data}(x)dx + \int_x p_{data}(x)(\log 2 + \log \frac{p_{data}(x)}{p_G(x) + p_{data}(x)} \\ &+ p_G(x)(\log 2 + \log \frac{p_G(x)}{p_G(x) + p_{data}(x)} dx \\ &= -\log 2.2 + \int_x p_{data}(x)(\log \frac{2.p_{data}(x)}{p_G(x) + p_{data}(x)} \\ &+ \int_x p_G(x)(\log \frac{2.p_G(x)}{p_G(x) + p_{data}(x)} dx \end{aligned}$$

$$C(G) = -\log 4 + \int_x p_{data}(x)(\log \frac{2.p_{data}(x)}{\frac{p_G(x)+p_{data}(x)}{2}} + \int_x p_G(x)(\log \frac{p_G(x)}{\frac{p_G(x)+p_{data}(x)}{2}} dx \tag{11}$$

Recall from Kullback-Leibler divergence [1],

$$D_{KL}(P \parallel Q) = \int p(x) \log \frac{p(x)}{q(x)} \tag{12}$$

Kullback-Leibler divergence is the measure of how much a given distribution differs from a second distribution. Using this identity in equation (11),

$$C(G) = -\log 4 + KL(p_{data} \parallel \frac{p_G(x) + p_{data}(x)}{2}) + KL(p_G \parallel \frac{p_G(x) + p_{data}(x)}{2}) \tag{13}$$

It is to be noted that KL divergence is always non-negative, therefore $-\log 4$ is the global minimum of $C(G)$. Moreover, by definition, Jensen-Shanon divergence [4] between two distributions is defined as,

$$JSD(P \parallel Q) = \frac{1}{2} D_{KL}(P \parallel M) + \frac{1}{2} D_{KL}(Q \parallel M)$$

Using this identity in equation (10),

$$C(G) = -\log 4 + 2.JSD(p_{data} \parallel p_G) \tag{14}$$

From definition of the Jensen-Shanon divergence, $JSD(p_{data} \parallel p_G)$ is only 0 when $p_{data} = p_G$ which leads to the conclusion that global minimum of $C(G)$ is $-\log 4$.

3 Additional Materials

Theorem 1 (Radon-Nikodym) *Suppose Ω is a non-empty set and $\mathcal{A}$ a σ-field on it. Suppose μ and ν are σ-finite measures on $(\Omega, \mathcal{A})$ such that for all A $\in \mathcal{A}$, $\mu(A) = 0$ implies $\nu(A) = 0$. Then,*
1. There exists $z : \Omega \to [0, \infty)$ measurable such that for all A in$\mathcal{A}$, $\nu(A) = \int_A z d\mu$.
2. Such a z is unique upto a.e. equality (with respect to μ)
3. z is integrable with respect to μ if and only if ν is a finite measure.

Theorem 2 (Kullback-Leibler Divergence) *Let $p(x)$ and $q(x)$ are two probability distributions of a discrete random variable x. That is, both $p(x)$ and $q(x)$ sum up to 1, and $p(x) > 0$ and $q(x) > 0$ for any x in X. Then, $D_{KL}(p(x), q(x))$ is defined as,*

$$D_{KL}(P \parallel Q) = \int p(x) \log \frac{p(x)}{q(x)}$$

KL-divergence is not symmetrical. That is, $D_{KL}(P \parallel Q) \neq D_{KL}(Q \parallel p)$. D_{KL} drops to 0 for area where $p(x) \to 0$. For example, in the figure 2, the red curve *corresponds to $D(p, q)$ and it drops to zero when $x > 2$ where p approaches 0. The KL-divergence $DL(p, q)$ penalizes the generator if it misses some modes of images: the penalty is high where $p(x) > 0$ but $q(x) \to 0$. Nevertheless, it is acceptable that some images do not look real. The penalty is low when $p(x) \to 0$ but $q(x) > 0$ (Poorer quality but more diverse samples). On the other hand, the reverse KL-divergence $DL(q, p)$ penalizes the generator if the images does not look real: high penalty if $p(x) \to 0$ but $q(x) > 0$. But it explores less variety: low penalty if $q(x) \to 0$ but $p(x) > 0$ (Better quality but less diverse samples).*

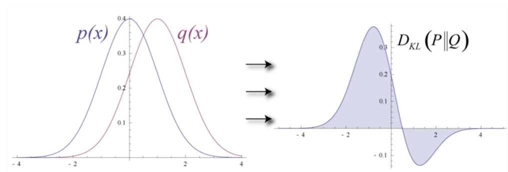

Figure 2: KL Divergence

Theorem 3 (Jensen-Shanon Divergence) *Jensen–Shannon divergence (JSD) is a symmetrized, finite and smoothed version of Kullback–Leibler divergence which is defined as*

$$JSD(P \parallel Q) = \frac{1}{2} D_{KL}(P \parallel M) + \frac{1}{2} D_{KL}(Q \parallel M)$$

where $M = \frac{P+Q}{2}$.

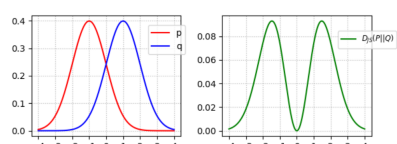

Figure 3: JS Divergence

In probability theory and statistics, JSD is applied as a popular method to measure the similarity between two probability distributions. It is also known as information radius (IRad) or total divergence to the average. The JSD is bounded by 1, given that one uses the base 2 logarithm $0 \leq JSD(P \parallel Q) \leq 1$. For Napierian logarithm, which is commonly used in statistical mechanics, the upper bound is $\ln 2 : 0 \leq JSD(P \parallel Q) \leq \ln 2$. Figure 3 shows the symmetry of JS divergence.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[3] Jonathan Hui. Gan—why it is so hard to train generative adversarial networks! https://medium.com/@jonathan$_hui/gan-why-it-is-so-hard-to-train-generative-advisory-networks-819a86b3750b, jun2018$.

[4] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

[5] Steven Lord, Fedor Sukochev, et al. Measure theory in noncommutative spaces. *SIGMA. Symmetry, Integrability and Geometry: Methods and Applications*, 6:072, 2010.

[6] Scott Rome. An annotated proof of generative adversarial networks with implementation notes. https://srome.github.io/An-Annotated-Proof-of-Generative-Adversarial-Networks-with-Implementation-Notes/, mar 2019.