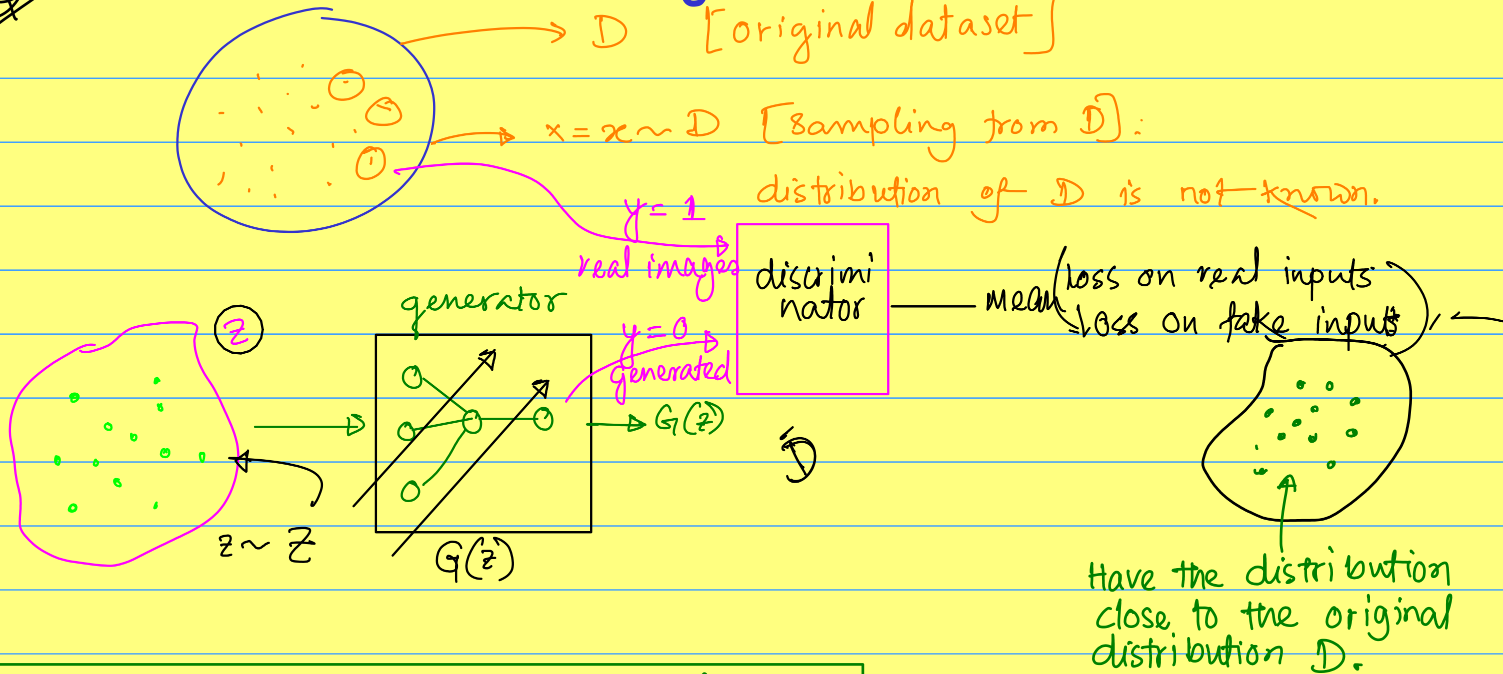
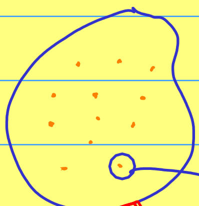
Understanding loss Function of GAN.



Basic Convensions to understand loss function:



$p_{data}(x)$: original data distribution (unknown to us / But we want to know).

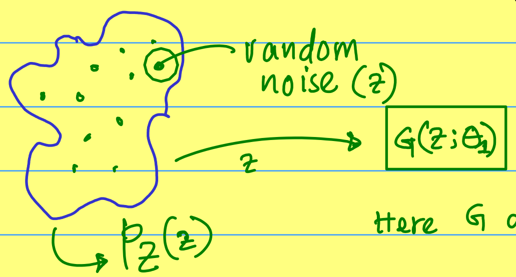
• : $X = x$ [single instantiation of the data from the real data.]

labels $y=1$

$D(x;\theta_2)$ → $D(x)$ transformed | probability that x is coming from original dataset

$D(G(z)) \in [0,1]$ Here $D(x) \in [0,1]$ Basic probability.

feeding it to discriminator



$p_g(x)$: we are trying to make $p_{data}(x) \approx p_g(x)$

Here G and D are differentiable function.

Loss function (Binary Cross Entropy)

$$L(\hat{y}, y) = [y \log \hat{y} + (1-y) \log(1-\hat{y})]$$

Loss contribution from real images:

Label for the original data coming from $p_{data}(x)$ will be $y = 1$ and for the output produced by the discriminator $D(x)$ the labels will be $\hat{y} = D(x) = 1$ "

So the amount of loss from the real images will be

$$L(\hat{y}, y) = L(D(x), 1) = [1 \cdot \log D(x) + 0 * \log(1 - D(x))] = \boxed{\log(D(x))} — (A)$$

Loss contribution from the generated images:

For data coming from the generator $G(z)$, the ground truth is $y=0$ and prediction $\hat{y} = D(G(z))$. Therefore

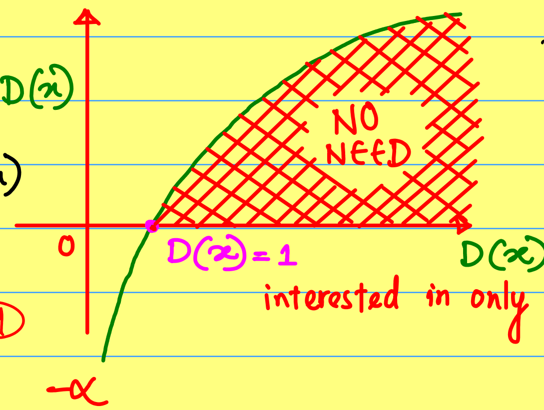
$$L(D(G(z)), 0) = \left[(1-0)\log\left(1 - D(G(z))\right)\right]$$
$$= \log\left(1 - D(G(z))\right) \text{ ——— } B$$

Objective of the Discriminator: We want to maximize both the A and B

A $\left[\log(D(x))\right]$

B $\left[\log\left(1 - D(G(z))\right)\right]$

log plot:



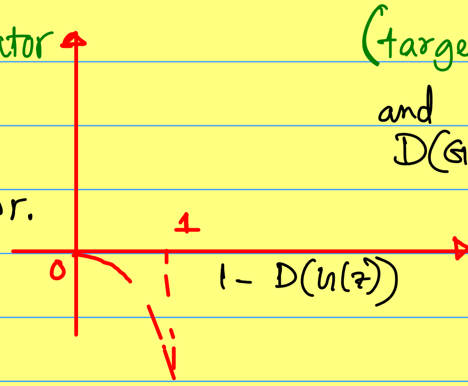
Therfore the objective function

$$\max\left\{\log(D(x)) + \log\left(1 - D(G(z))\right)\right\} = D$$

Objective of the generator: fool the discriminator

$D(G(z)) \to 1$ [tentative move]

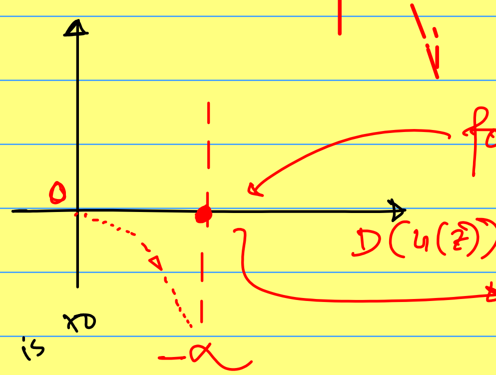
(target of discriminator) and force $D(G(z)) \to 0$



→ Not at all dependent on the generator.

Look at the second term.

$$\log\left(1 - D(G(z))\right)$$



fooling point

if $D(G(z)) = 1$ then it is fooled.

at this point $\log(1-1) = -\infty$

very small value compared to the other things.

So here ~~for generator~~ the aim is to

$$\min\left[\log(D(x)) + \log\left(1 - D(G(z))\right)\right]$$

there is no gen term here.

→ produce data to the original data distribution.

Therefore combining both of the optimization task we get

$$\min_G \max_D \left\{ \log(D(x)) + \log(1 - D(G(z))) \right\} \quad \text{— } C$$

Important:

(#) when generator is considered (minimized) only the $\log(1-D(G(z)))$ term will be considered. i.e only the parameters of $G(z)$ will be updated.

(#) when the discriminator is optimized (maximized), then both of the term will be considered. i.e Both parameters of D and G will be updated.

.

The above equation C is only for a single value or point. But for all the points in the dataset of n points. it will be something like

$$\min_G \max_D \frac{1}{m}\sum_{i=1}^{m} \log(D(x_i)) + \frac{1}{m}\sum_{i=1}^{m} \log(1 - D(G(z_i)))$$

which can be re written as the expectation value

$$\boxed{\min_G \max_D V(D,G) = \underbrace{\left\{ E_{x \sim P_{data}(x)} \log(D(x)) + E_{z \sim P_z(z)} \log(1 - D(G(z))) \right\}}_{\min_G \max_D}}$$

Lecture : 29

Intutive understanding of loss function

Finding the best discriminator :

Proposition: For a fixed G, the optimal discriminator D is

$$D_G^*(x) = \frac{P_{data}(x)}{P_{data}(x) + P_g(x)}$$ optimal / best.

Proof: the training criterion for the discriminator D given any generator G is to maximize the equation (A) which is

$$\min_G \max_D V(D,G) = \min_G \max_D \left\{ E_{x \sim P_{data}(x)} \log(D(x)) + E_{z \sim P_z(z)} [\log(1 - D(G(z)))] \right\}$$

if G is given, that means the generator is given i.e. fixed, then the goal is to maximize the remaining descriminator.

Hence the optimal discriminator is defined as

$$D_G^* = \underset{D}{\operatorname{argmax}} \; V(D,G)$$

Here please note that $E_{p(x)}[x] = \int_x x \, p_x(x) \, dx$ ——— (1)

value

prob of that value x in $P_x(x)$.

Expectation of the variable x having the probability density function $p(x)$.

Therefore the optimal discriminator will be

$$D_G^* = \underset{D}{\operatorname{argmax}} \left\{ E_{x \sim P_{data}(x)} \log(D(x)) + E_{z \sim P_z(z)} [\log(1 - D(G(z)))] \right\}$$

$$= \underset{D}{\operatorname{argmax}} \{ V(D, G) \} \text{ ——— (2)}$$

what value of D will maximize $V(D,G)$.

Expectation of any random variable x with probability density function pdf of $P_x(x)$ is denoted as

$$E_{p(x)}[x] = \int_x x \cdot P_x(x) \, dx \text{ ——— (1)}$$

Putting the equation (1) in the equation (2) we get

$$V(G,D) = \int_x P_{data}(x) \cdot \log(D(x)) \, dx + \int_z P_z(z) \log(1 - D(G(z))) \, dz$$

But the original paper directly jumped from the above equation to the following equation

$$V(G,D) = \int_x P_x(x) \log D(x)\,dx + \int_x P_g(x) \cdot \log(1 - D(x))\,dx .$$

Therefore we need to find out how the following is possible

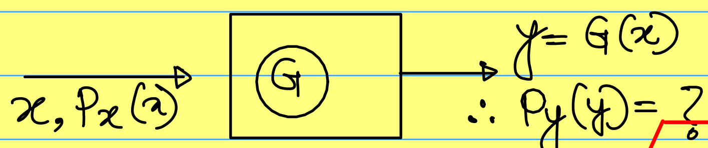
$$\boxed{\int_z P_z(z) \log D(G(z))\,dz = \int_x P_g(x) \cdot \log(1 - D(x))\,dx}$$

Probability Concept: If the probability density function (PDF) of a random variable x is given as $P_x(x)$, it is possible to calculate the probability density function of some other variable $y = G(x)$. This is called "change of variable" and is defined as

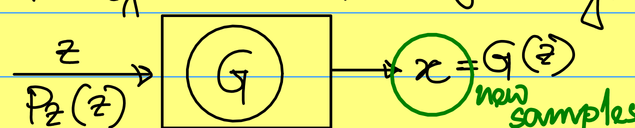
$$\boxed{P_y(y) = P_x(G^{-1}(y)) \left| \frac{d}{dy}(G^{-1}(y)) \right|}$$

important.

Here $y = G(x) = x^2$
$= x^2 + 4$
$= x^3 + x^2 + 1$
it can be any function.



Now in our case the generator takes the input z. Therefore the system is like the following.



Therefore,

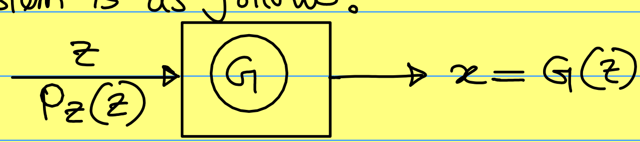
$$P_g(x) = P_z(G^{-1}(x)) \cdot \frac{d}{dx}(G^{-1}(x)).$$

our case

We need to show that

$$\int_z P_z(z) \cdot \log(1 - D(G(z)))\,dz = \int_x P_g(x) \log(1 - D(x))\,dx$$

Since here the system is as follows:



$x = G(z)$
$z = G^{-1}(x).$

Now if we assume that G is invertible, then it is true that

$$z = G^{-1}(x) \quad [\text{inverse of a function}]$$

Therefore from the left hand side we get

$$\int_z p_z(z) \cdot \log(1 - D(G(z)))\, dz$$

$$= \int_x p_z(G^{-1}(x)) \log(1 - D(x))\, dG^{-1}(x)$$

$$= \int_x p_z(G^{-1}(x)) \cdot \log(1 - D(x)) \frac{dG^{-1}(x)}{dx} dx.$$

using the relation $p_g(x) = p_g(G^{-1}(x)) \cdot \frac{d}{dx} G^{-1}(x)$. [derived earlier].

$$= \int_x p_z(G^{-1}(z)) \cdot \log(1 - D(x)) \cdot \frac{d\, G^{-1}(x)}{dx} dx$$

$$= \int_x \underbrace{\boxed{p_z(G^{-1}(z)) \cdot \frac{d}{dx} G^{-1}(x).}}_{p_g(x)} \log(1 - D(x))\, dx$$

$$= \int_x p_g(x) \cdot \log(1 - D(x))\, dx \quad \text{———} \quad (3)$$

Now the whole loss function looks like the following:

$$V(G, D) = \int_x p_{data}(x) \cdot \log D(x)\, dx + \int_z p_z(z) \cdot \log(1 - D(G(z)))\, dz$$

$$= \int_x p_{data}(x) \cdot \log D(x)\, dx + \int_x p_g(x) \cdot \log(1 - D(x))\, dx$$

$$= \int_x \left[p_{data}(x) \cdot \log(D(x)) + p_g(x) \cdot \log(1 - D(x)) \right] dx$$

Thus the optimal D^* for a given G is obtained by maximizing $V(G, D)$ from above expression.

So we will find the maximum value of the integrand and choose the value at the maximum value to be the optimal value of D for a given G i.e. D^*_G. Therefore,

$$\frac{d}{dD(x)} = \left[p_{data}(x) \log(D(x)) + p_g(x) \log(1 - D(x)) \right] = 0$$

But here $\frac{d}{dD(x)} = \frac{P_{data}(x)}{D(x)} - \frac{P_g(x)}{1-D(x)} = 0.$ $\quad D(x) \rightarrow D^*_G$

Formula: $\boxed{\frac{d}{dx}\log(x) = \frac{1}{x}}$

$$\therefore \frac{P_{data}(x)}{D(x)} = \frac{P_g(x)}{1-D(x)}$$

$$\Rightarrow P_{data}(x).(1-D(x)) = P_g(x).D(x).$$

$$\Rightarrow P_{data}(x) - P_{data}(x).D(x) = P_g(x).D(x).$$

$$\Rightarrow P_{data}(x) = P_{data}(x)D(x) + P_g(x)D(x).$$

$$\Rightarrow D(x).(P_{data}(x) + P_g(x)) = P_{data}(x).$$

$$\Rightarrow D^*_G(x) = \frac{P_{data}(x)}{P_{data}(x) + P_g(x)}. \quad \text{(Proved)}$$

To prove that its the maximum value we differentiate the following equation again $\frac{P_{data}(x)}{D(x)} - \frac{P_g(x)}{1-D(x)} = 0$

$$\Rightarrow \frac{-P_{data}(x)}{D(x)^2} - \frac{P_g(x)}{(1-D(x))^2} < 0$$

Hence it was in fact the maximum value of the function.

Therefore the optimal D for a given G will be

$$\boxed{D^*_G(x) = \frac{P_{data}(x)}{P_{data}(x) + P_g(x)}}.$$

→ optimal.

→ maximal.

It is theoretically solvable but practically impossible to solve because the apriory $P_{data}(x)$ is not possible to calculate deterministically.

The role of the generator is to reverse that of the discriminator. i.e. minimizing the objective function. So the optimal G that minimize the loss function occurs when $D = D^*_G$. Therefore we get the optimal G^* as

$$G^* = \arg\min_G V(D^*_G, G)$$

At this point we must show that the optimization problem stated in (A) has unique solution G^* and this solution satisfies the $P_g = P_{data}$.

So from previous proof $D^*_G = \dfrac{P_{data}(x)}{P_{data}(x) + P_g(x)}$

Therefore substituting the optimal value of D^*_G in the equation (A) we get

$$G^* = \arg\min_G V(D^*_G, G)$$

$$= \arg\min_G \int_x \left[P_{data}(x) . \log\left(D^*_G(x)\right) + P_g(x) . \log\left(1 - D^*_G(x)\right) \right] dx$$

Now again we have $D^*_G = \dfrac{P_{data}(x)}{P_{data}(x) + P_g(x)}$; Substituting this we get

$$G^* = \arg\min_G \int_x \left[P_{data}(x) . \log\left(\frac{P_{data}(x)}{P_{data}(x) + P_g(x)}\right) + P_g(x) . \log\left(1 - \frac{P_{data}(x)}{P_{data}(x) + P_g(x)}\right) \right] dx$$

$$= \arg\min_G \int_x \left[P_{data}(x) . \log\left(\frac{P_{data}(x)}{P_{data}(x) + P_g(x)}\right) + P_g(x) . \log\left(\frac{P_{data}(x) + P_g(x) - P_{data}(x)}{P_{data}(x) + P_g(x)}\right) \right] dx$$

$$= \arg\min_G \int_x \left[P_{data}(x) . \log\left(\frac{P_{data}(x)}{P_{data}(x) + P_g(x)}\right) + P_g(x) . \log\left(\frac{P_g(x)}{P_{data}(x) + P_g(x)}\right) \right] dx$$

Now we add and subtract $(\log 2) P_{data}(x)$ and $(\log 2) P_g(x)$ in the above equation, then we get

$$G^* = \arg\min_G \int_x \left[(\log 2 - \log 2) P_{data}(x) + P_{data}(x) \log \frac{P_{data}(x)}{P_{data}(x) + P_g(x)} + (\log 2 - \log 2) P_g(x) + P_g(x) \log \frac{P_g(x)}{P_{data}(x) + P_g(x)} \right] dx$$

$$G^* = \arg\min_G \int_x \Big[-\log 2 \left(P_{data}(x) + P_g(x)\right) + P_{data}(x) \left\{ \log 2 + \log\left(\frac{P_{data}(x)}{P_{data}(x) + P_g(x)}\right) \right\}$$

$$+ P_g(x) \cdot \left\{ \log \frac{P_g(x)}{P_{data}(x) + P_g(x)} + \log 2 \right\} \Big] dx$$

$$G^* = \arg\min_G \log 2 \int_x \left(P_{data}(x) + P_g(x)\right) dx$$

$$+ \int_x P_{data}(x) \left[\log 2 + \log\left(\frac{P_{data}(x)}{P_{data}(x) + P_g(x)}\right) \right] dx$$

$$+ \int_x P_g(x) \left[\log 2 + \log\left(\frac{P_g(x)}{P_{data}(x) + P_g(x)}\right) \right] dx$$

Now for any PDF its integral is always 1. Therefore

$$\int_x P_{data}(x)\, dx = 1$$

and $\int_x P_g(x)\, dx = 1$.

$$G^* = \arg\min_G \log 2 \,(1+1)$$

$$+ \int_x P_{data}(x) \cdot \log\left(\frac{P_{data}(x)}{\left(\frac{P_{data}(x) + P_g(x)}{2}\right)}\right) dx$$

$$+ \int_x P_g(x) \cdot \log\left(\frac{P_g(x)}{\left(\frac{P_{data}(x) + P_g(x)}{2}\right)}\right) dx$$

From KL divergence equation we get

$$KL(P(x) \| Q(x)) = \int_x P(x) \cdot \log \frac{P(x)}{Q(x)}\, dx.$$

$$\therefore G^* = \arg\min_G \left\{ -\log 4 + KL\left(P_{data}(x) \,\Big\|\, \frac{P_{data}(x) + P_g(x)}{2}\right) + KL\left(P_g(x) \,\Big\|\, \frac{P_{data}(x) + P_g(x)}{2}\right) \right\}$$

one thing to look for here is that there is no term G on the right hand side to optimize. But we have P_g which is actually dependent on G.

Previously we used $P_z(G^{-1}(z))$ $\qquad x = G(z)$

$G^{-1}(z) = x.$

$P_g(x)$

JSD = Jensen Shannon Divergence.

$$G^* = \arg\min_G \left\{ -\log 4 + 2\,JSD\left(P_{data}(x) \,\|\, P_g(x)\right) \right\}$$

Here we get

$$G^* = \underset{G}{\arg\min} \left\{ -\log 4 + 2\,JSD\left(P_{data}(x) \,||\, P_g(x)\right) \right\}$$

where $JSD\left(P(x) \,||\, Q(x)\right) = \frac{1}{2}\left[KL\left(P(x) \,||\, M(x)\right) + KL\left(Q(x) \,||\, M(x)\right) \right]$

where $M(x) = \dfrac{P(x) + Q(x)}{2}$

But the JSD term becomes 0 when $P_g(x) = P_{data}(x)$. which is our ultimate goal but it is very hard to achieve.

Therefore in the above equation JSD is 0 only when $P_g(x) = P_{data}(x)$. which minimizes the argument and the value obtained is $-\log 4$.

$$G^* \quad \underset{G}{\arg\min} \left\{ -\log 4 + \underbrace{2\,JSD\left(P_g(x) \,||\, P_{data}(x)\right)} \right\}$$

one way to make it zero is to make $P_g(x) = P_{data}(x)$. the optimization process is forcing this to happen.

Theorem: The global minimum of the criterion $G^* = \underset{G}{\arg\min}\, V(D_G^*, G)$ is achieved if and only if $P_g(x) = P_{data}(x)$. At that point G^* will have the value of $-\log 4$

Proof: From previous proof we know $D_G^* = \dfrac{P_{data}(x)}{P_{data}(x) + P_g(x)}$

Now if we put $P_g(x) = P_{data}(x)$ then

$$D_G^* = 1/2$$

Now we put it in the expression of $G^* = \underset{G}{\arg\min}\, V(D_G^*, G)$

$$\therefore G^* = \underset{G}{\arg\min}\, V(1/2, G)$$

$$= \underset{G}{\arg\min} \left\{ \int_x \left[P_{data}(x) . \log\left(D(x)\right) + P_g(x) . \log\left(1 - D(x)\right) \right] dx \right.$$

$$= \underset{G}{\arg\min} \left\{ \int_x \left[P_{data}(x) \log(1/2) + P_g(x) . \log(1/2) \right] dx \right.$$

$$= \arg\min_{G} -\log(2)\left[\cdot \int_x P_{data}(x)\,dx + \int_x P_g(x)\,dx\right]$$

$$= -\log(2) * \left[1 + 1\right]$$

$$= -\log 4$$

$$\int_x P_{data}(x)\,dx = 1 \quad [\text{PDF}]$$

$$\int_x P_g(x)\,dx = 1 \quad [\text{PDF}].$$