Module 22.3 : Generative Adversarial Networks - The Math Behind it

- We will now delve a bit deeper into the objective function used by GANs and see what it implies
- Suppose we denote the true data distribution by $p_{data}(x)$ and the distribution of the data generated by the model as $p_G(x)$
- What do we wish should happen at the end of training?

$$p_G(x) = p_{data}(x)$$

- Can we prove this formally even though the model is not explicitly computing this density?
- We will try to prove this over the next few slides

Theorem

The global minimum of the virtual training criterion $C(G) = \underset{D}{max}\, V(G, D)$ is achieved **if and only if** $p_G = p_{data}$

is equivalent to

Theorem

1. **If** $p_G = p_{data}$ then the global minimum of the virtual training criterion $C(G) = \underset{D}{max}\, V(G, D)$ is achieved **and**
2. The global minimum of the virtual training criterion $C(G) = \underset{D}{max}\, V(G, D)$ is achieved **only if** $p_G = p_{data}$

Outline of the Proof

The 'if' part: The global minimum of the virtual training criterion $C(G) = \max_D V(G, D)$ is achieved **if** $p_G = p_{data}$

(a) Find the value of $V(D, G)$ when the generator is optimal *i.e.*, when $p_G = p_{data}$

(b) Find the value of $V(D, G)$ for other values of the generator *i.e.*, for any p_G such that $p_G \neq p_{data}$

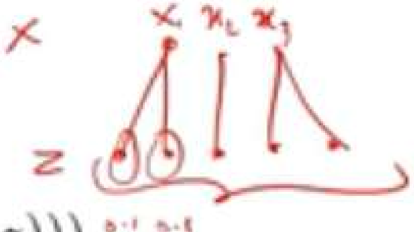
(c) Show that $a < b \ \forall \ p_G \neq p_{data}$(and hence the minimum $V(D, G)$ is achieved when $p_G = p_{data}$)

The 'only if' part: The global minimum of the virtual training criterion $C(G) = \max_D V(G, D)$ is achieved **only if** $p_G = p_{data}$

- Show that when $V(D, G)$ is minimum then $p_G = p_{data}$

- First let us look at the objective function again

$$\min_{\phi} \max_{\theta} \quad [\mathbb{E}_{x \sim p_{data}} \log D_\theta(x) + \mathbb{E}_{z \sim p(z)} \log(1 - D_\theta(G_\phi(z)))]$$



- We will expand it to its integral form

$$\min_{\phi} \max_{\theta} \int_x p_{data}(x) \log D_\theta(x) + \int_z p(z) \log(1 - D_\theta(G_\phi(z)))$$

- Let $p_G(X)$ denote the distribution of the X's generated by the generator and since X is a function of z we can replace the second integral as shown below

$$\min_{\phi} \max_{\theta} \int_x p_{data}(x) \log D_\theta(x) + \int_x p_G(x) \log(1 - D_\theta(x))$$

- The above replacement follows from the *law of the unconscious statistician* (click to link of wikipedia page)

- Okay, so our revised objective is given by

$$\min_{\phi} \max_{\theta} \int_x \left(p_{data}(x) \log D_\theta(x) + p_G(x) \log(1 - D_\theta(x))\right) dx$$

- Given a generator G, we are interested in finding the optimum discriminator D which will maximize the above objective function
- The above objective will be maximized when the quantity inside the integral is maximized $\forall x$
- To find the optima we will take the derivative of the term inside the integral w.r.t. D and set it to zero

$$\frac{d}{d(D_\theta(x))} \left(p_{data}(x) \log D_\theta(x) + p_G(x) \log(1 - D_\theta(x))\right) = 0$$

$$p_{data}(x) \frac{1}{D_\theta(x)} + p_G(x) \frac{1}{1 - D_\theta(x)}(-1) = 0$$

$$\frac{p_{data}(x)}{D_\theta(x)} = \frac{p_G(x)}{1 - D_\theta(x)}$$

$$(p_{data}(x))(1 - D_\theta(x)) = (p_G(x))(D_\theta(x))$$

$$D_\theta(x) = \frac{p_{data}(x)}{p_G(x) + p_{data}(x)}$$

- This means for any given generator

$$D_G^*(G(x)) = \frac{p_{data}(x)}{p_{data}(x) + p_G(x)}$$

- Now the if part of the theorem says "if $p_G = p_{data}$ ...."
- So let us substitute $p_G = p_{data}$ into $D_G^*(G(x))$ and see what happens to the loss functions

$$D_G^* = \frac{p_{data}}{p_{data} + p_G} = \frac{1}{2}$$

$$\begin{aligned} V(G, D_G^*) &= \int_x p_{data}(x) \log D(x) + p_G(x) \log\left(1 - D(x)\right) dx \\ &= \int_x p_{data}(x) \log \frac{1}{2} + p_G(x) \log \left(1 - \frac{1}{2}\right) dx \\ &= \log 2 \int_x p_G(x) dx - \log 2 \int_x p_{data}(x) dx \\ &= -2 \log 2 \qquad = -\log 4 \end{aligned}$$

- So what we have proved so far is that if the generator is optimal ($p_G = p_{data}$) the discriminator's loss value is $-\log 4$
- We still haven't proved that this is the minima
- For example, it is possible that for some $p_G \neq p_{data}$, the discriminator's loss value is lower than $-\log 4$
- To show that the discriminator achieves its lowest value "if $p_G = p_{data}$", we need to show that for all other values of p_G the discriminator's loss value is greater than $-\log 4$

- To show this we will get rid of the assumption that $p_G = p_{data}$

$$
\begin{aligned}
C(G) &= \int_x \left[p_{data}(x) \log \left(\frac{p_{data}(x)}{p_G(x) + p_{data}(x)} \right) + p_G(x) \log \left(1 - \frac{p_{data}(x)}{p_G(x) + p_{data}(x)} \right) \right] dx \\
&= \int_x \left[p_{data}(x) \log \left(\frac{p_{data}(x)}{p_G(x) + p_{data}(x)} \right) + p_G(x) \log \left(\frac{p_G(x)}{p_G(x) + p_{data}(x)} \right) + (\log 2 - \log 2)(p_{data} + p_G) \right] dx \\
&= -\log 2 \int_x (p_G(x) + p_{data}(x))\, dx \\
&\quad + \int_x \left[p_{data}(x) \left(\log 2 + \log \left(\frac{p_{data}(x)}{p_G(x) + p_{data}(x)} \right) \right) + p_G(x) \left(\log 2 + \log \left(\frac{p_G(x)}{P p_G(x) + p_{data}(x)} \right) \right) \right] dx \\
&= -\log 2 (1 + 1) \\
&\quad + \int_x \left[p_{data}(x) \log \left(\frac{p_{data}(x)}{\frac{p_G(x) + p_{data}(x)}{2}} \right) + p_G(x) \log \left(\frac{p_G(x)}{\frac{p_G(x) + p_{data}(x)}{2}} \right) \right] dx \\
&= -\log 4 + KL \left(p_{data} \| \frac{p_G(x) + p_{data}(x)}{2} \right) + KL \left(p_G \| \frac{p_G(x) + p_{data}(x)}{2} \right)
\end{aligned}
$$

- Okay, so we have

$$C(G) = -\log 4 + KL\left(p_{data}||\frac{p_{data} + p_g}{2}\right) + KL\left(p_G||\frac{p_{data} + p_G}{2}\right)$$

- We know that KL divergence is always ≥ 0

$$\therefore C(G) \geq -\log 4$$

- Hence the minimum possible value of $C(G)$ is $-\log 4$

- Now let's look at the other part of the theorem
  If the global minimum of the virtual training criterion $C(G) = \max_{D} V(G, D)$ is achieved then $p_G = p_{data}$
- We know that

$$C(G) = -\log 4 + KL\left(p_{data} \| \frac{p_{data} + p_g}{2}\right) + KL\left(p_G \| \frac{p_{data} + p_G}{2}\right)$$

- If the global minima is achieved then $C(G) = -\log 4$ which implies that

$$KL\left(p_{data} \| \frac{p_{data} + p_g}{2}\right) + KL\left(p_G \| \frac{p_{data} + p_G}{2}\right) = 0$$

- This will happen only when $p_G = p_{data}$ (you can prove this easily)
- In fact $KL\left(p_{data} \| \frac{p_{data}+p_g}{2}\right) + KL\left(p_G \| \frac{p_{data}+p_G}{2}\right)$ is the Jenson-Shannon divergence between p_G and p_{data}

$$KL\left(p_{data} \| \frac{p_{data} + p_g}{2}\right) + KL\left(p_G \| \frac{p_{data} + p_G}{2}\right) = JSD(p_{data} \| p_G)$$

which is minimum only when $p_G = p_{data}$

21:49 / 22:29