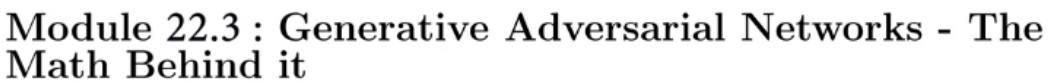
Lecture – 22.3
Generative Adversarial Networks – The Math Behind it

Refer Slide Time: (00:13)



So, so far what we have seen now, I'm going back in history and back in the past and starting as a scientist finished model 23.2. So, so far what we have seen is that an intuitive explanation of GANs, you had this generator, a discriminator you had this minimax objective function which you minimize with respect to the parameter of the generator and maximize with respect to the parameters of the discriminator,

Refer Slide Time: (00:36)

Architecture guidelines for stable Deep Convolutional GANs

- Replace any pooling layers with strided convolutions (discriminator) and fractional-strided convolutions (generator).
- Use batchnorm in both the generator and the discriminator.
- Remove fully connected hidden layers for deeper architectures.
- Use ReLU activation in generator for all layers except for the output, which uses tanh.
- Use LeakyReLU activation in the discriminator for all layers

and then we saw this over all algorithm

Refer Slide Time: (00:37)

- We will now look at one of the popular neural networks used for the generator and discriminator (Deep Convolutional GANs)
- For discriminator, any CNN based classifier with 1 class (real) at the output can be used (e.g. VGG, ResNet, etc.)

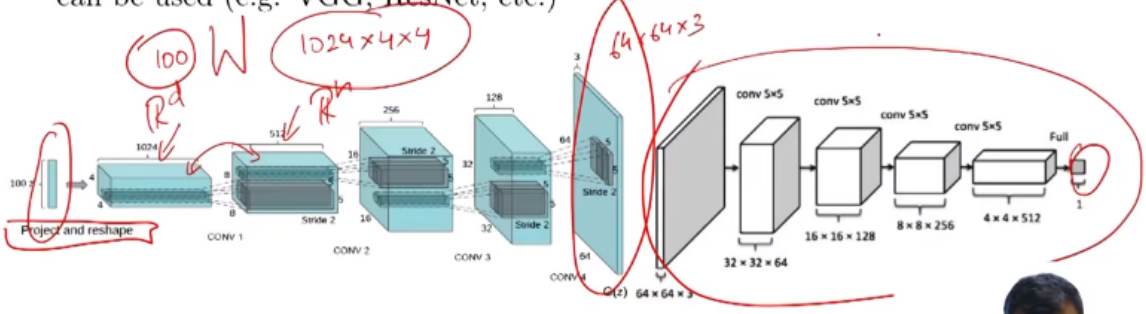


Figure: Generator (Redford et al 2015) (left) and discriminator (Yeh et al used in DCGAN

for training

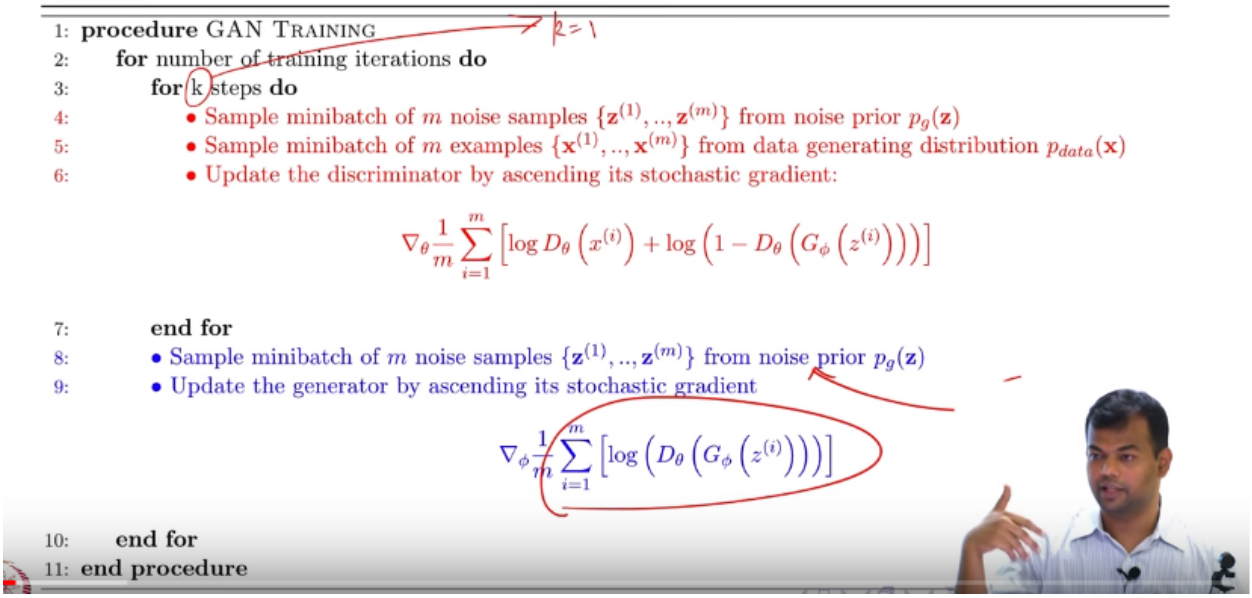
Refer Slide Time: (00:37)

Module 22.3 : Generative Adversarial Networks - The Math Behind it

them,

Refer Slide Time: (00:38)

With that we are now ready to see the full algorithm for training GANs



which alternates between the generated parameters and the discriminative parameters,

Refer Slide Time: (00:42)

Module 23.2: Generative Adversarial Networks - Architecture

Okay?

Refer Slide Time: (00:43)

- We will now look at one of the popular neural networks used for the generator and discriminator (Deep Convolutional GANs)
- For discriminator, any CNN based classifier with 1 class (real) at the output can be used (e.g. VGG, ResNet, etc.)

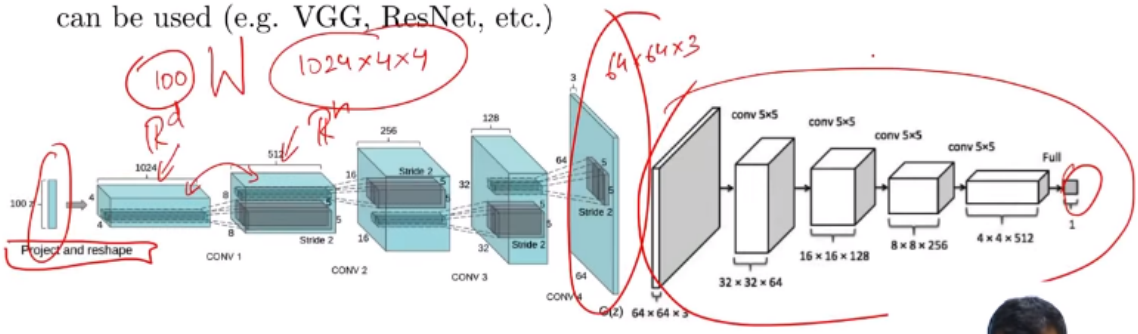


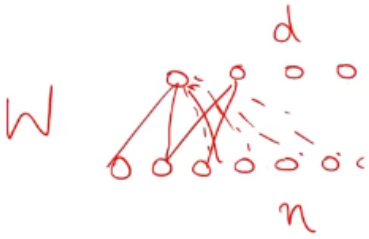
Figure: Generator (Redford et al 2015) (left) and discriminator (Yeh et al used in DCGAN

Now,

Refer Slide Time: (00:43)

Architecture guidelines for stable Deep Convolutional GANs

- Replace any pooling layers with strided convolutions (discriminator) and fractional-strided convolutions (generator).



Refer Slide Time: (00:44)

Architecture guidelines for stable Deep Convolutional GANs

- Replace any pooling layers with strided convolutions (discriminator) and fractional-strided convolutions (generator).
- Use batchnorm in both the generator and the discriminator.

we want to look at some math behind this intuition. Right? So, all we have done is you have set up an objective function and just assume all of that will work, we are not even defined, what does it mean by saying at the end the GAN, that we trained actually works what does it mean?

Refer Slide Time: (01:01)

- 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?

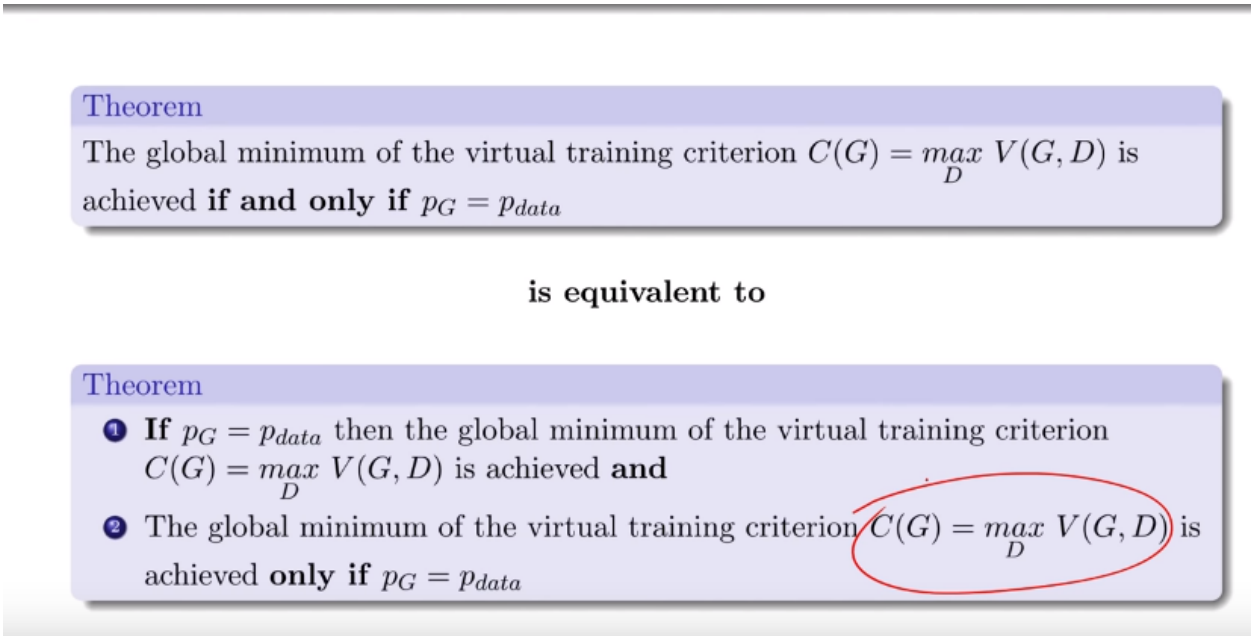
$$p_G(x) = p_{data}(x)$$

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

So, let's start with that, Right? So, we'll delve a bit deeper into this objective function and see what it implies? So, suppose you have this true data distribution, Right? So, you are given this training samples, Right? And from these training samples these are all the amnest images, each block is 1 amnest image, now these training samples and this samples actually come from some distribution. Right? On the other hand you have this generator which is also generating images and I can say all of these images are coming from the another distribution which is p of G, x, both these distributions are on top of x, where x is your images 1024 dimensional or n dimensional. So, both these distributions are over the same set of random variables, but one is the true distribution, one is the generated distribution. Okay? If everything works fine, what do you want to happen at the end, pG, x, is equal to p data x. Right? That's what we mean

when we say that it should work. Okay? But is this the object, if that was the case we should have said this as the objective function, Right? we should have said that minimize the KL divergence between these two or whatever other probability distance function that you know how? Right? But we didn't do that we use some other cryptic objective function which depended on some cross entropy or some score and so on, Right? So, what we need to show is that and the reason we did that is because we did not have a pG, x, Right? We did not explicitly model pG of x, unlike Auto regressive models or RBMS or VAEs there was no p of x in this generator, Right? We never came up with the formula for p of x, that's why we couldn't have that as the objective function, all you can hope for is that whatever this pseudo objective function, we have set up that eventually leads to this condition, even though we have explicitly not computed p of x or p of Gx, does that make sense, is it. Okay? What we need to actually show, is that clear nothing. Right? So, So, what you want to say is that. So, here's let me state it this way you want to show that if the discriminative loss function, is at its minimum value, can I guarantee that that can happen only if pG is equal to P data, is that fine, is that a good statement. Okay? So, we'll try to prove that, Okay?

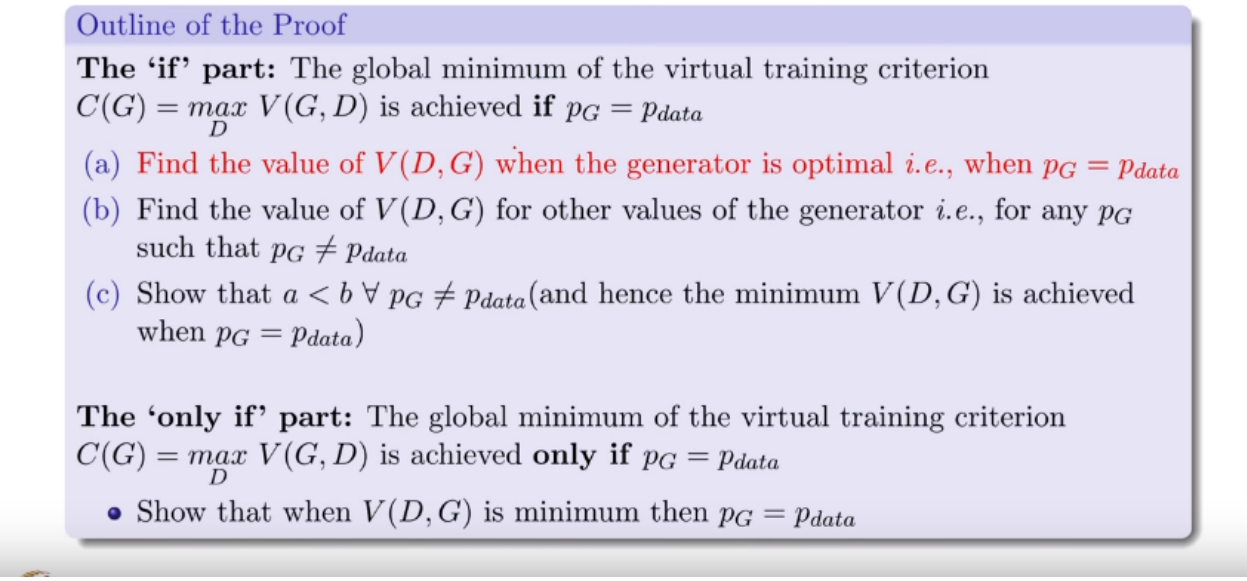
Refer Slide Time: (03:17)



So, here's this theorem statement. Right? This is from the paper, that the global minimum of this training objective rate and this max V of G, D, is nothing but that two part objective function which I had written it. So, this V of G, D, is just that two-part objective function which I had written where you have an expectation over the true samples and an expectation over the generated samples, Right? and I am trying to maximize this with respect to the discriminator and I have to show that. So, actually yeah! So, here's the theorem and I'll just explain what it means. So, remember that the goal of the generator is to minimize

the maximum of this value, wait it's a first you have the discriminator you compute some loss function with respect to that and the goal of the generator is to minimize the maximum value of that, is that fine? Because it's minimize or maximize, is that fine, everyone is. Okay. So, now the theorem states that this minimum will be achieved if and only if pG is equal to p data, Right? So, what it means is that whatever objective function I have setup, if I'm able to achieve that objective function then, I can be sure that pG is equal to P data, which would have actually been my true objective function, is that fine, is a statement of the proof clear, if I prove this then, you'll be fine, that I mean we have gone back to our original goal and proved it, Okay? Now, any theorem which has an if and only if part can always be split into these two parts Right? the if part and the only part. So, they've part is that if pG is equal to P data then, the global minimum of the virtual criteria would be achieved. Right? And the only if part is that if this is achieved then pG has to be equal to P data, there is no other way that could have been achieved you get the difference between the if and only if part. So, if, if pG is equal to P data then, this will go to its minimum that's fine, but if only if part is the reverse of this which means that if I know that this has gone to this minimum value then, I can be sure that PG has to be equal to P data. Right? So, I have to prove, prove the F part as well as the only if part. Okay?

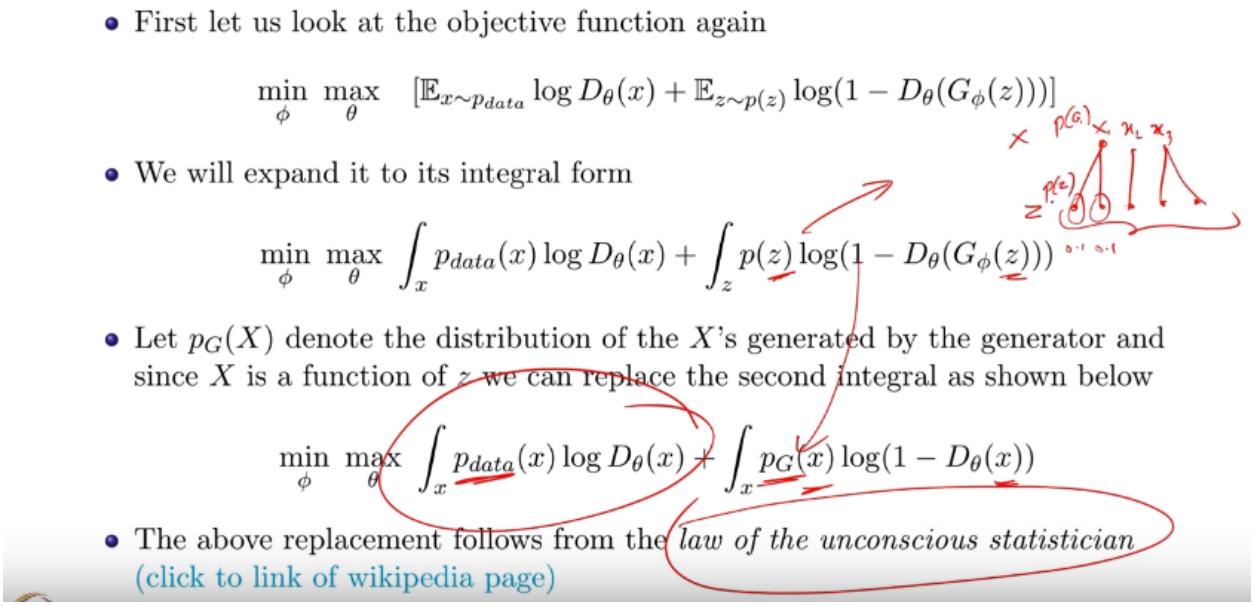
Refer Slide Time: (05:25)



So, here's the outline of the proof, we first look at the if part and then we look at the only if part. Okay? So, the if part, to show the if part I will show the following first, I have to find this value of D, G when PG is equal to P data, because that's what the if condition says. Okay. So, I'll find what this value of V, D, G is or rather actually C, G is. Okay? Then I'll find the same value when PG is not equal to P data, Okay?

So, I know what the values when PG is equal to data, I know what the values when PG not equal to P data, for any pG, any other pG. Now, what do I have to prove, to prove the if part, first one is always less than equal to. Right? Minima, is that fine! So, when PG equal to P data whatever value you get that's always less than equal to the value, when you get when PG is not equal to P data, is that fine? So, that's the third step that I need to prove that show that is less than equal to less than B or less than equal to, is also fine, is that. Okay? Everyone gets this three parts, I'll show what the value is, when P is equal to P data I'll show what the value is when PG not equal to P data and then, I will show that A is less than equal to B, if I can show that then, I can show that the minimize are chained only when PG is equal to P data, is the outline of the proof clear, does that make sense, anyone, who does not get that, please raise your hands, if you get this. Okay. Now, the only if part is that when this minimum is achieved, I need to show that PG is equal to P data that means, if someone tells me that I have achieved the minimum, then it has to be the case that PG is equal to P data, is that fine, the if part and the only if part. Okay? So, let's start with the if part and we look at the first step of the if part.

Refer Slide Time: (07:18)



So, this is the objective function, now what I'm going to do is, I'm going to replace the expectations by their integrals. Okay? And now I'm going to observe that first of all, observe that this is nothing but x, this is the generated x. Right? So, I know that this G, Phi of Z, is a function of Z. So, by this some cryptic rule actually, I can replace the second integral which was over Z by an integral over corresponding x is, everyone Okay? With this, the first integral has been copied as it is, but the second integral has been replaced by a different integral instead of Z. Now, I have an integral over x, why does this make sense,

what is the rule that I've used here? I've used change of variables that's a valid answer, but there's a slight problem there when you do change of variables, you have to assume that the function is invertible, we don't know whether G is invertible or not. Right? Okay? So, I'll give you an interview to explain intuitive explanation for why this makes sense, this actually comes from something known as the law of the unconscious statistician, I don't know who named it that way, but I'll give you an intuitive explanation for why this is Okay? Right? So, we'll try to understand, what this integral tried to tries to do and then say that it's, Okay. If I replace it by this integral, Okay? So, the first integral is actually over Z. So, these are all the Z's that you have. Okay? Now, given as Z is G, Phi of Z, a deterministic function or a random function? A deterministic function. I'm going to pass it through the neural network for a given Z, I am always going to get the same X, no matter how many times I pass it, Okay? So, from the Z domain to the X domain what kind of a function do I have, many to many, one to many, one to one or many to one, can two different Z's give me the same X, in two different sets give me the same X, that's possible. Think of a simple neural network with just, just classification. Right? you could give it to Apple images for both of them, it can give you the same output. Right? So, for two different inputs I can get the same output, but for the same input, can I get two different outputs. No. So, then what kind of a function is this, many to one. Okay? So, there could be many Z's here which correspond to the same X here. Okay? Is that fine? Now, this integral is actually over the Z's, I am integrating over these Z's based on their individual probabilities, is that fine? But now if I look at the X's their probabilities are actually just a function of these two because these if I just sum up these two Z properties I will get the probability of the given corresponding X, does that make sense, if all of these are say point one, point one. Okay? And only these two Z's result in this X and what's the probability of that X going to be, point two. Does that make sense, Okay? So, now instead of summing over these Z's, I can just sum over the X's and replace the probability of Z by the probability of X, Right? because whenever summing over the Z's, I was summing over these Phi terms but it turns out that these Phi term just collapse two three terms in the X domain. So, instead of summing over these Phi terms, I can sum over those three terms and replace them by the corresponding probabilities, does that make sense, please raise your hands, if it does? So, that's the intuitive explanation for why this works and there's actually a law behind, that you can go and prove it more formally, but you understand the intuition, that's fine. I don't care if you don't really know this law, because even I don't know it's Okay? So, that's how you will replace the Z's here by X's here and the reason I am doing that is I have an, already have one integral with respect to X, I wanted the second interval also, to be with respect to X and the second thing was I wanted P data and PG. So, I now have P data and PG, does that make sense, Okay? And remember this PG is actually a function of p, Z. You can write it as that, is it Okay? Fine.

Refer Slide Time: (11:43)

- 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

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

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

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

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

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

So, now this is our revised objective function. Okay? Now, when will this, achieve the inner thing, Right? if I look at this when will it achieve its maximal, maxima, when for every given X, the term inside the integral is maximized. Right? So, instead of integral, just think of it as a sum. So, you have a sum of some terms, you want to find the maximum of this sum. So, the sum would be maximized when every term in the sum is maximize, does that make sense. Okay? So, then I just look at the quantity inside the integral and I'll take its derivative with respect to D and set it to zero, Okay? So, I will take the derivative with respect to D and set it to zero, Okay? Can you help me in working this derivative, what will be the first step? P data upon 1 by period upon D, X into 1, plus PG over. Okay? I'll just write it is that fine, Okay? Now, I can just do some simple I'll take one term on one side. Okay? And I want an expression for D theta X, completely sure, it looks like it matters but I am not completely sure of the implications of that, Right? whether it matters to this achieving its final objective or not I am not sure that I need to work out Okay? So, we will come back to that let's not confuse the others about that. Okay? So, for now just see that this thing Right? So, you get the optimal discriminator, what is this D theta of X, is it a distribution, is it a score, what is it? It is score. So, what does it mean, that I will give you an X, the score, the best score that the discrimination assign, it is just take P data of X and divided by PG of X plus P data of X, Right? that's the optimal discriminator, is that fine. Okay?

Refer Slide Time: (13:44)

- 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

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

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

Now, for any given generator, if I give you the generator, then the optimal discriminator is given by this, Okay? But what is the condition that we had in the if part of the theorem, P data is equal to PG. So, if I substitute that here what do I get, half, Right? So, that's why when you end training your discriminator should actually irrespective of whether it's a true image or a fake image it's so confused that it just assign the score of 0.5 to both, I can't really distinguish. So, I'll just say with half probability that this is real or half cloudy that this is fake, Okay? That makes intuitive sense, Okay? So, now this, this quantity, Right? that we were interested in or rather the max of this quantity, I'll just substitute the max for each of these guys inside that's the one which we have computed here. So, I'll substitute that, and this is what I get, Okay? Is that fine, what is this integral, what is this integral, one, Right? integral over power density function, what is this integral? One. So, what's the final output answer that you get, oh wait! So, this should be plus, all I think this should be minus. So, what you'll get is minus 2 log 2 which is minus log 4. So, this is the value you achieve, when PG is equal to P data. So, I have computed the value that you will achieve when P is equal to P data that was the first part or the first step of my proof, is that fine? Now, what do I need to do when pG not equal to P data what happens Right?

Refer Slide Time: (15:32)

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 $p_G = p_{data}$

So, that's why we will go to the second part, we have done the first part. Now, you look at the second part.

Refer Slide Time: (15:36)

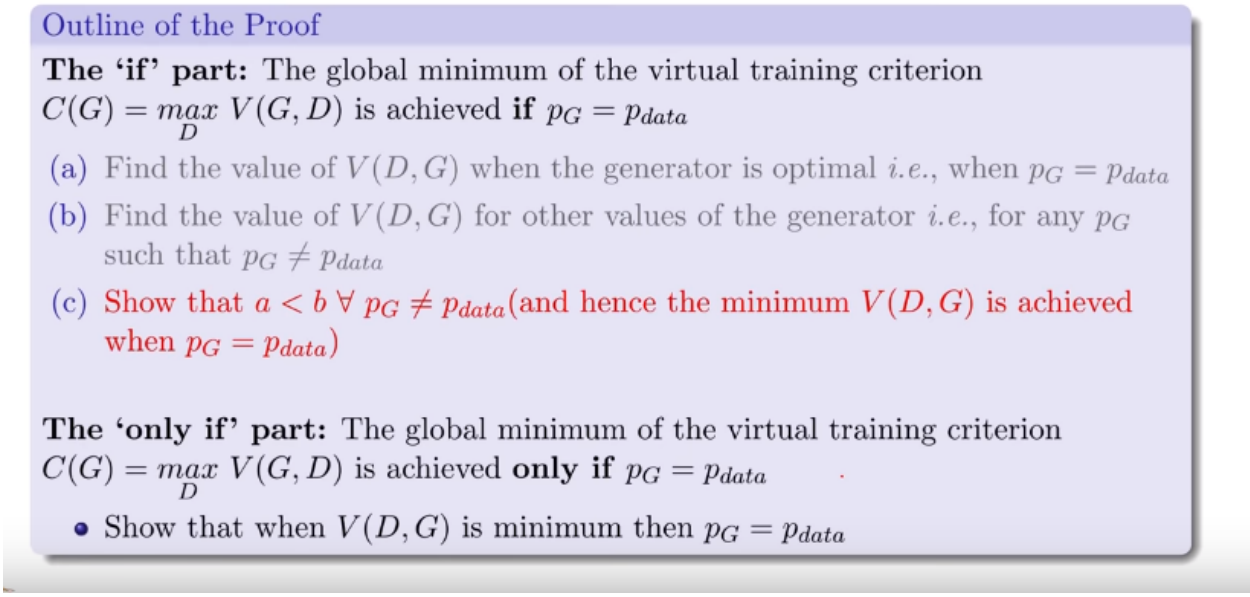
- To show this we will get rid of the assumption that $p_G = p_{data}$

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

So, we will throw away this assumption, that PG is equal to P data, Okay? So, you look at PG not equal to P data. So, let's see what happens in that case. So, I am not going to substitute half here, because half is only when P is equal to P data. So, I'm going to work with when PG is not equal to P data, Okay? So, now since I know that some where I need to compare with a log 4. So, this is some trickery that I'm going to do, I'm going to add this term and you see that what I'm adding is just a zero, Right? because it's log two minus log two, I'm just adding a zero. So, this is a fair thing, I'm only doing this because I know where I need to reach and just trying some manipulation. So, that I reach where I want to reach, Okay? So, this is

not something which you can come up with this is because you know what the answer is. So, you have added to this, I'm sure all of you have done similar proofs and different high school courses also, Okay? Is that fine? I've added this effectively; I've just added a zero. Now, what I'm going to do is, I'm just going to rearrange some terms here and there, Okay? So, first note that I can take this minus log two and this part has one integral, Okay? Is that Okay? And then each of these other log two is I am going to split it between the remaining two integrals, don't try to really squint and try to understand this, this is very very simple, I genuinely mean that is just some rearrangement of terms and once you do all this, Right? what you will end up with, Okay? Let's not go there. So, you will end up with minus log 4, plus some quantity, can you tell me what this quantity is, first what is this? It's a distribution, what about this distribution, what does this term look like? KL divergence between P data and the term inside the bracket and what about this KL divergence between PG and the term inside the bracket, everyone gets that, Right? So, you can just think of this as P and this is Q. So, you have integral P log or, Sorry! P log P by Q and integral Q log Q by P or whatever, Right? So, you can just write it as sum KL divergence, Okay? So, it's actually the KL divergence between P data and this other distribution, is it Okay? So, you have minus log 4 plus 2, K L divergences, Okay? So, when PG is equal to P data this value is minus log 4, when I don't assume PG is equal to period, I get minus log 4 plus 2, KL divergences. So, given two, these two statements, what can you tell me, what do you know about KL divergence always greater than equal to 0. Now, can you tell me something, Okay. So, let's look at the steps.

Refer Slide Time: (18:36)



So, now at this point what I have done is, I have done the step two of the theorem of the proof, I have shown you what is the value, when PG is not equal to P data, is that fine,

Refer Slide Time: (18:45)

- 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 $-\log 4$
- But this is the value that $C(G)$ achieves when $p_G = p_{data}$ (and this is exactly what we wanted to prove)
- We have, thus, proved the **if part** of the theorem

Okay? Now, what I need to show you is that the minimum is attained only when PG is equal to P data, Okay. So, the general term is this, when for any PG any p data, Okay? Now, what do we know about the KL divergence greater than equal to 0. So, what then, what can you say about CG, it's always going to be greater than equal to minus log 4, that means minus log 4 is the lower bound, it's the minimum value that CG can take, and when did CG take that value, when P is equal to P data. So, what are which proved now, we have proved the if part, if PG is equal to P data then CG takes its minimum value, is that fine? Everyone gets that, Okay? So, that's the end of the if part.

Refer Slide Time: (19:31)

- 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 $p_G = p_{data}$

Now, for the KL or the only if part what I need to show is that if CG has taken its minimum value someone has told me here's the CG and it has achieved its minimum value then, it has to be the case that PG is equal to P data, Okay? Now, what's the minimum value of CG, minus log 4, we know that, Okay? So, let's start with that. So, the general formula is minus log 4 plus this. Now, someone has told me that CG has attain edits minimum value that means, what do I know CG is equal to minus log 4, that implies that two other terms are 0. So, KL divergence of this, and KL divergence of this is equal to 0, Okay? So, you can now prove that since KL divergence can only be greater than equal to zero. So, these two terms cannot cancel each other. So, they both have to go to zero and you can easily prove that this will happen only if PG is equal to P data, in fact you don't even need to prove that whatever I have written here is actually the symmetric version of KL divergence and it is known as the Jenson Shannon divergence, Okay? So, this is actually the Jenson Shannon divergence between P data and PG and the Jensen Shannon divergence would be zero, only when P data is equal to PG. Okay? So, whichever you want to prove it, you can prove it either prove that this is equal to zero, I mean this equal to 0, implies that P is equal to P data or just assume that this is equal to Jensen Shannon divergence not assume that that's the actual relation and that will be 0, only when P is equal to P data, Okay? So, have you proved the, if and the only if part? So, what have we effectively proved that whatever objective function we have chosen, this minimax problem that we had chosen actually optimizing that that is the same as optimizing this other objective function, which is PG should be equal to P data, that's what we have proved, this guy will attain attained its optimum value only when PG is equal to P data, is that fine? Right? So, that's the more formal proof about, why the particular objective function for GANS. Okay? So, with that I finally end the lecture and eCos. So, thank you and I hope you guys enjoyed it and I hope you guys learned a few things and I

hope you guys do well in the end sem, and on the remaining assignments also. So, this end sem will again we have the same flavor as quiz 1, it will not be of the same flavor as quiz 2. Okay? Thank you.