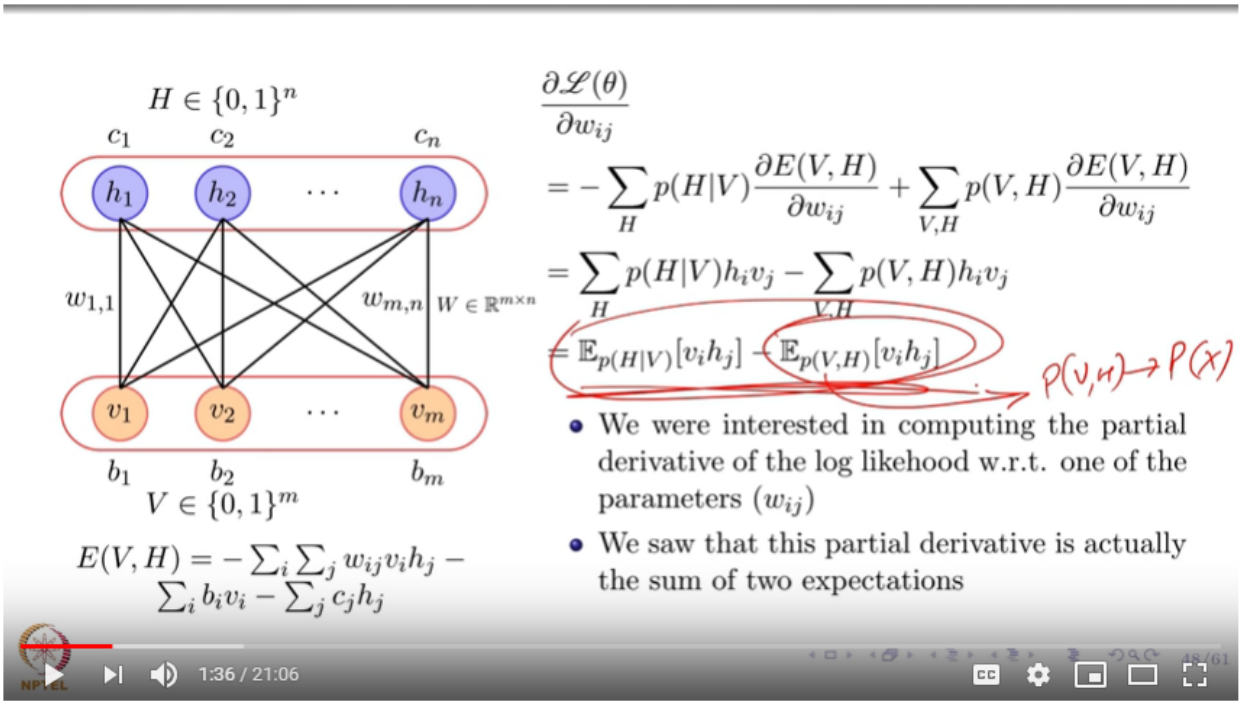
Lecture 19.4
Training RBMs Using Gibbs Sampling

Refer slide time:(0:13)

- Okay, so we are now ready to write the full algorithm for training RBMs using Gibbs Sampling
- We will first quickly revisit the expectations that we wanted to compute and write a simplified expression for them

So, now with this background, we are now ready to write the full algorithm for training, RBMs using Gibbs sampling. So, before that we'll just do a quick recap of, what was that quantity which was irritating, us and now whether we have been able to take care of that quantity. Right?

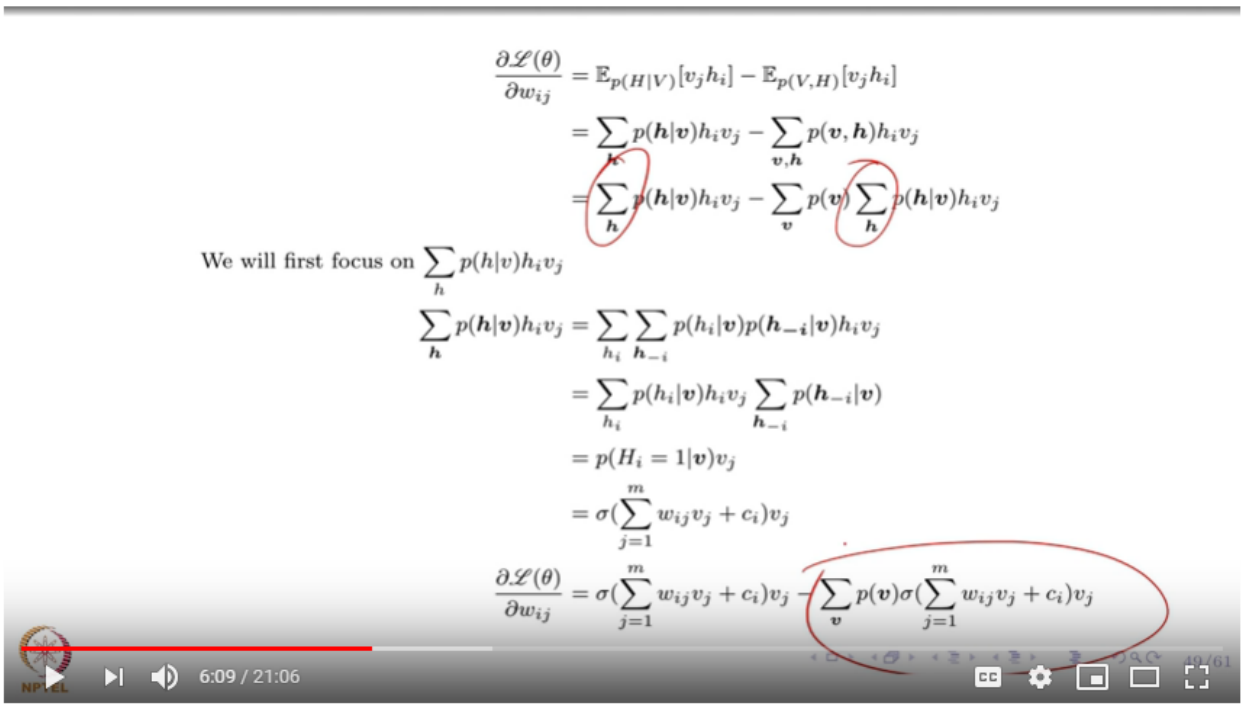
Refer slide time:(0:29)



So, what we were interested, in as usual is the gradient of the loss function, with respect to the some parameter, the parameter was W IJ and we did some trickery and we figured out that, this is actually a

sum of two expectations. Right? And our problem was that we cannot compute this expectation, because we cannot draw samples from this Joint Distribution, I have written it as P of V comma H, but is the same as P of X, in the previous discussion. But now with this Gibbs sampling procedure, I have given you away of sampling from this distribution or in other words, I have given your way of approximating this expectation interact able expectation, by an empirical expectation. Right? So, that's why all this is connected, that this is where we started this was our root cause of all the problems that we had this interact able expectation and we wanted to do something about this you, still need to complete the full story, but we will do that first thing that we are going to do is they're going to look at these two expectations, in their true form, that means in their exponential form and then try to simplify it and then arrive at something, which then will try to deal with, with the help of Gibson. Okay?

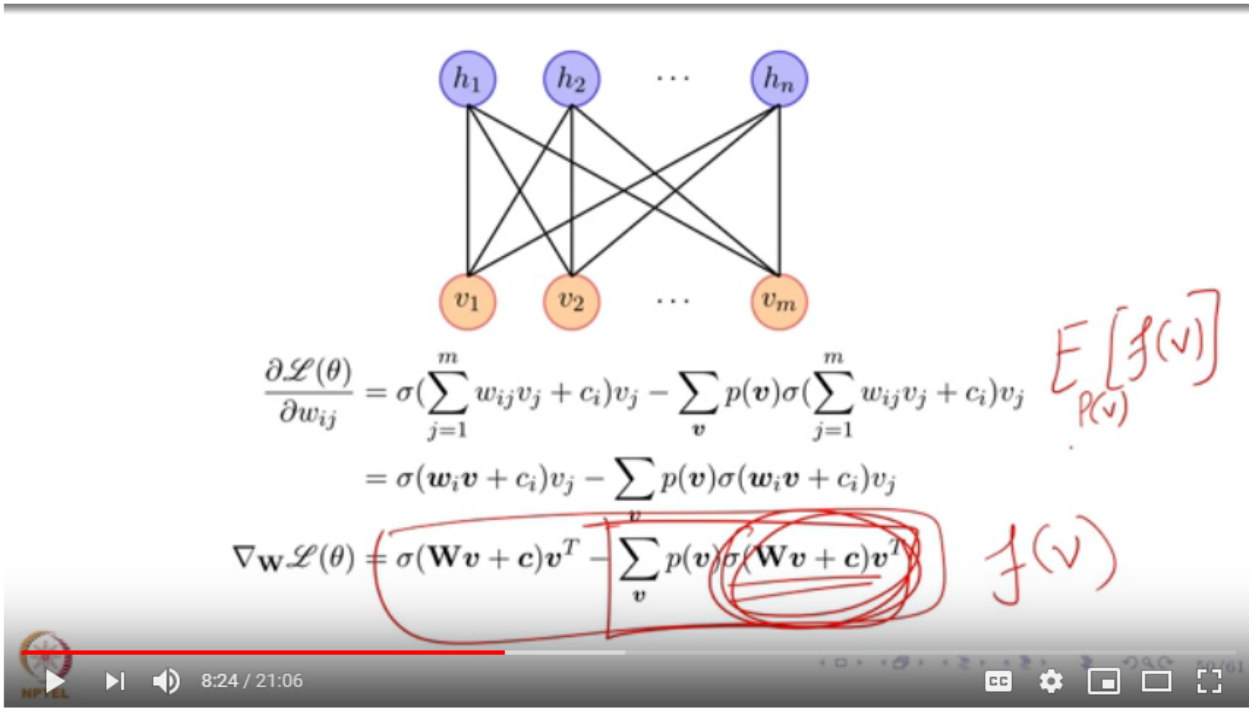
Refer slide time:(01:41)



So, I look at some very simple mats. So, you have to take my word for this that this is what these two expectations were we had actually, rigorously derived them and this is the function that we had inside the expectation. Okay? So, I can write this as following this is the true expectation, I sum over all possible values of H and my expectation is over the distribution, P of H given V and this is the function that I care about HIV J. Okay? And the second quantity, is actually a sum over all possible values of V comma H, under the distribution P of e comma H and this is the function that I care about, is that. Okay? Now let's just split this into two parts. Right? So, this is a summation over V and a summation over H and this P of V comma H, I have just splitted P of V, into P of H given V, is that. Okay? I just separated out the two summations, every n fine at this point, please. Okay? Now you see that this quantity actually repeats at both the cases, in both the terms. Okay? Let's look at that quantity and see if we can simplify, it so, we will just focus on this quantity. Right? Now again this sum is actually a sum over H1, H 2, H 3, up to HM. So, I can split it into two sums, one is H I and the other is all the remaining a H is that fine, can you guys

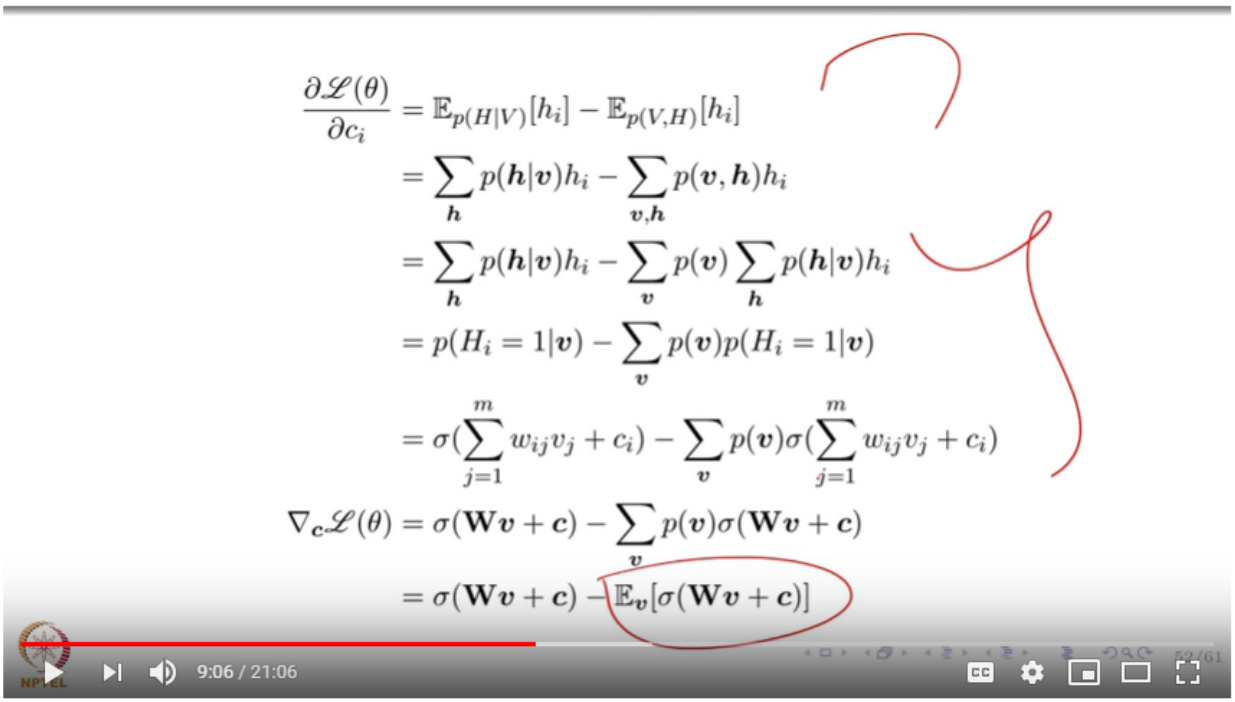
be show some more and two its straightforward. Right? I mean oh you are blank because, you are surprised by what simple stuff I am teaching or is it clear. Okay? So, this is one of the edges and these are all the remaining edges, hence this is bold and this is not bold. Okay? And ok again remember that this conditional can be factorized right that's what our assumption was that this was P of H 1, H 2, H 3, given V but we can just write it as P of H 1, given V into all the other guys given V that's the factorization that we have assumed in the case of, RBMs everyone okay with that. Okay? So, now I'll just take out the terms, which are dependent on H I, outside and I'll keep this inside. What is this summation? If you can answer this I will know that this point in continuing this lecture otherwise there is no point in continuous, what is the summation everyone, everyone write the summation of all possible values that this guy can take, so, that's going to be1. Okay? Is that fine. Okay? So, that means I will continue with the lecture. Okay? And what is this actually? This summation has how many terms in our case? How many terms is this summation half? Which are the two terms? H I equal to 0 and as I equal to 1. Right? So, this actually is P of H I equal to 0, given V into H IV J and P of H I equal to 1,given V into h I VJ, which of these two terms will disappear, H I equal to 0. Right? Because you have a product of 0here so, that will disappear so, this just boils down to P of H I ,equal to 1 given V into V J is that. Okay? So, notice that in the first term, which looked nasty it had this big summation over all possible values of H, it turns out that this is actually not nasty it just boils down to a very simple, computation. So, at least the first expectation, we have actually got rid of it straight away right we don't even need sampling to deal with, it, it just boils down to a very simple term, what about the second expectation, this that's again simplified. Right? So, it just becomes ok the first term is this. So, now if I substitute, this equal to this in this equation, I'll get the following. Okay? In particular note that from both the cases the summation over H has, disappeared. Right? But unfortunately in the second term, the summation over V, still remains. Okay? And this V is intractable, this summation is still intractable, because it's still 2 raise to M. Okay? Is it fine if you okay with this. Right?

Refer slide time:(06:14)



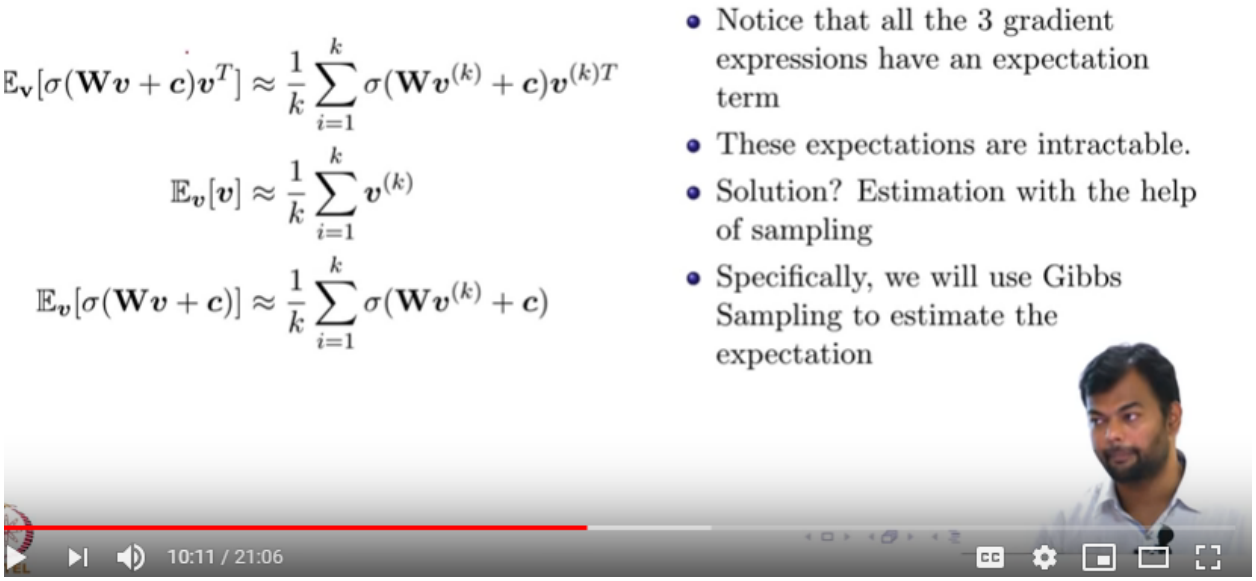
So, this is what we have come up with and now, can you give me a simple way of writing this term, what does this look like? If I assume that this is actually W. Right? This matrix of all the weights is W, then can you tell me a simplified form for this, oh capital W into V. What if this is not a matrix product? This is a there's a dot product. Right? So, what are the two vectors involved here, WI and V. Right? So, this is WI and V. so, ayat column of W and B, how many of you are fine with this piece raise your hands good. So, this is WI and V and the same simplification everywhere. Okay? Now this is the partial derivative of the loss function, with respect to one parameter, now can you give me the gradient of the loss function, with respect to the entire weight matrix, can you generalize from here, the hint is in this formula. Right? Now we have one column of the matrix, when we generalize we will have the entire matrix. Right? If you get that intuition the rest of it you can work out. Right? So, this is how the generalized form is going to look like. Okay? So, starting with the same recipe that we compute the gradient with respect to one of the parameters or rather the partial derivative, with respect to one of the parameters, we now have a formula for the gradient and this formula still has one problem, which is this, this is how can I write this actually? If I denote this as F, of V, how can I write this, what is this quantity actually? If I denote the sigmoid part as F of V, what is this expectation of F of V under P of V. Right? And this is still tractable intractable hence we need to do something ok fine is that fine. Okay?

Refer slide time :(08:28)



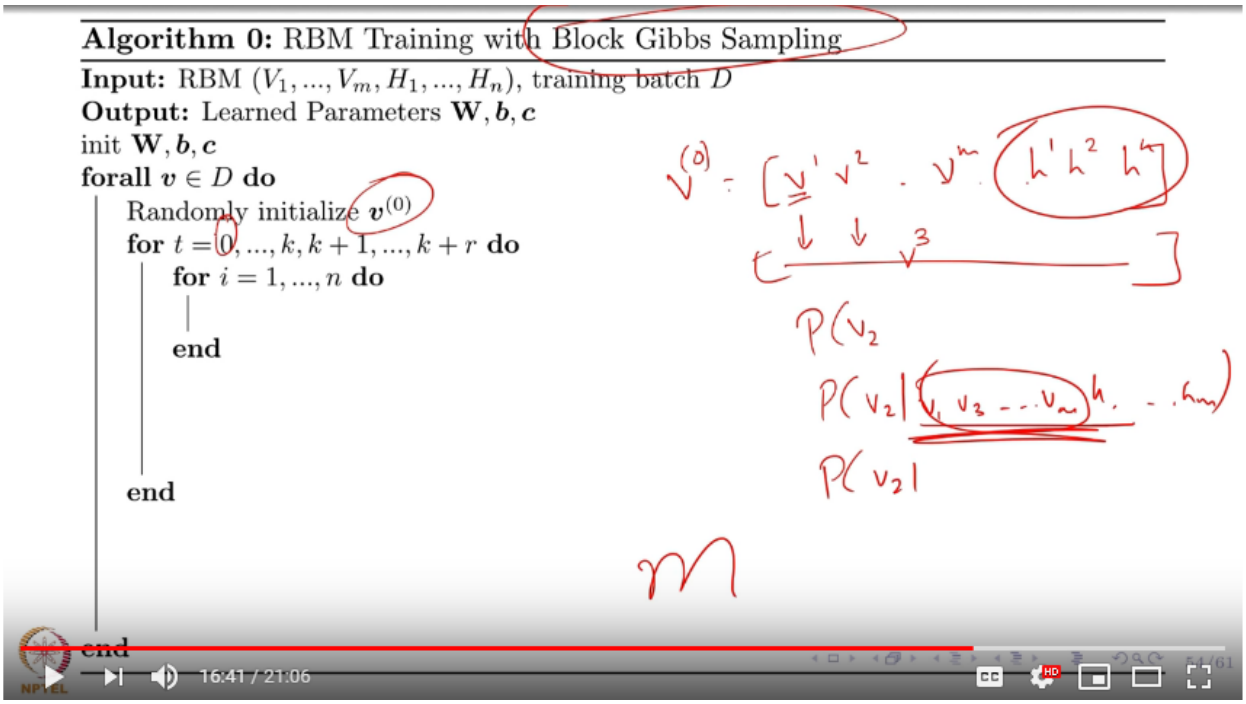
And that was W IJ, we can do the same computation for BJ that means the gradient with respect to BJ or the partial derivative with respect to one of the beasts, and then we can generalize to the gradient so, you can go back and check this, again this will have this, bad expectation term and the same thing holds for the C's, again the C's will have this bad expectation. Right? So, this math is very similar to the mat that we did for W IJ it's all straightforward you can go back and check it I'm not rushing, because I do not have time Ian ways would have rushed. Right? Okay? This is very straightforward based on all the stuff that we have done in this course before this should be very trivial to understand, the key thing to note here is that.

Refer slide time :(09:09)



All the three gradients that we have computed that is the gradient with respect to W, the gradient with respect to B and a gradient with respect to C, contains some expectation term. Okay? And this expectation with respect, to the random variable V, which is a collection of all the visible, variables these expectations are intractable. So, now to be able to do gradient descent, although we have a formula for the gradient, we need to, do something, what will we do we'll get rid of this interact able computation by doing give something. So, we will try to replace these expectations, by their empirical estimates and how will I get these case samples, by running the Markov chain wrong enough. So, that I start getting samples from the distribution that I care about, once that happens, I will take KSAT samples and substitute that here. And the reason I can do that is that because, this chain is easy to set up drawing samples from this chain is easy. So, all these expectations, I can replace by their empirical, estimate, by running Gibbs sampling everyone gets that. Okay?

Refer slide time :(10:16)

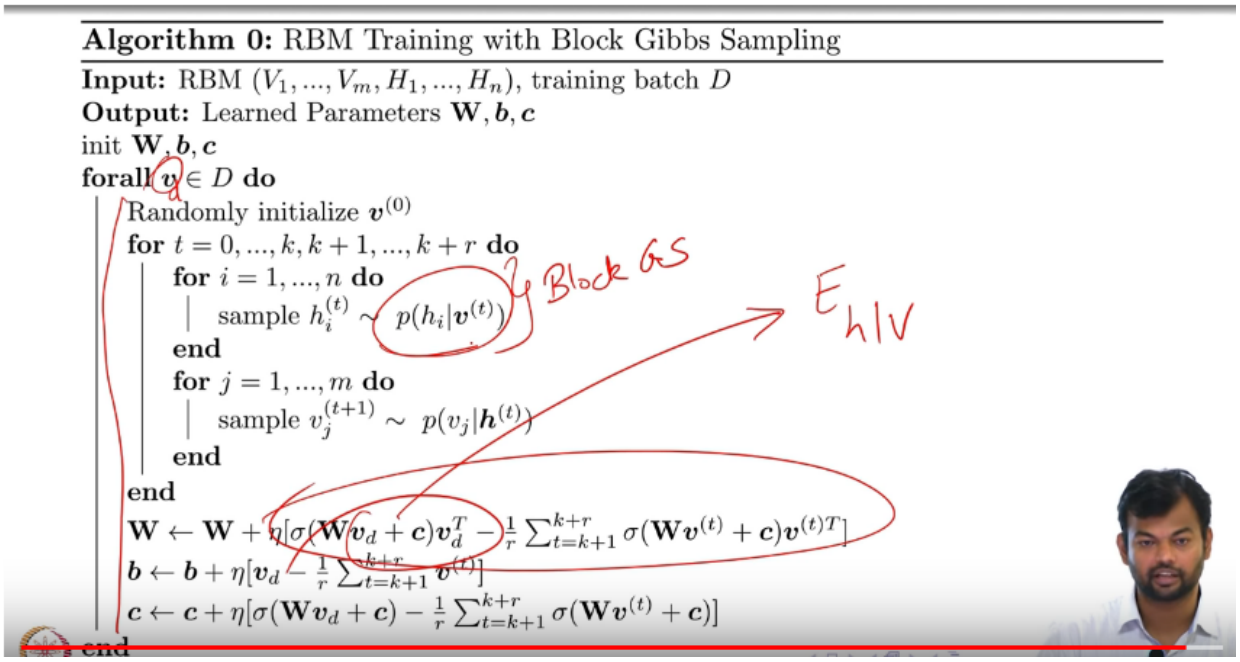


So, now let's actually, write the full RBM training algorithm, pay attention this is exactly, what you are going to implement. Okay? In the last assignment, which will be released for the weekend. So, as input you have this RBM, which has these visible and hidden variables. Okay? And just assume that this is the training data given to you, all the training data that you have, the output is the learned parameters WP and C, you'll start by initializing the WPS NC is using whatever favorite, initialization algorithm you have, not two zeros not two equal values not two large values. Now for all training examples belonging to your training data, what do you need to do. Okay? Without me telling you the algorithm. Okay? Tell me what the last three steps of this algorithm are going to be, what are the last three steps of this coloring would be W is equal to W minus ETA into gradient, B's we'll be minus ETA into gradient and C is equal to C my z10 to be again. Right? So, I need to first, the remaining steps are going to be about computing the gradient, for computing the gradient I need to run, Gibbs sampling. So, the first block in this empty space is going to be about, Gibbs sampling. Okay? So, I randomly initialize V0 this is the starting point of my of my Markov chain. So, I have picked up some V0 now I'm going to run the Markov chain for the large number of time steps. Okay? Why I split it as k and k plus 1 to R, I'm assuming that after Kit is going to, converge remember the proof that we have done does not tell us, when it will converge, it will only converge asymptotically. Right? Only as n tends to infinity in, practice the way you do it is that you run it for many, many, many time steps and after say some, 1 million time steps or 10 K time steps or hundred K time steps, you assume it has reached the stationary distribution and from that point onwards, whatever samples you get, you assume it comes from your distribution that's how you have to do it in practice, in practice you cannot do the asymptotic stuff which is up to infinity. Okay? Fine. So, that's how I split it, now at every time step, what do I do in the Gibbs chain? In the Markov chain, I have started with v 0 at

every time step what am I going to do so, remember that v0, is actually, a collection of all my random variables and in fact I can also add the edges here. Right? So, it's 2 up to H in and I have given up 2 VM and I actually have the values for these, these are zeros or ones or whatever. Right? Now what do I do in my mouth off chain, I need to pick one of these values, sample and you value for that, in practice we do something very simple, we don't have this uniform distribution, what we do is that we go of these random variables one by one in order, we'll first pick up the random variable one, change its value or retain its value based on the conditional distribution that we had defined earlier, then pick up the random variable tow, change its value or keep its value depend random on the distribution that we are defined earlier. So, what I am saying is that, our qi is very simple, it's just periodic, pick up first, pick up second, pick up third and so on just go over all the random variables, one by one, instead of sampling the value of five. Right? And the argument here or the rationale here is that, anyways you are going to run it for many, many time steps. So, anyways you are going to hit all of them once, so why not just go over them in order that's the simplified, thing so you don't have qi in practice, you just go over all of them once. Now let's look at this, even more carefully and this is the difference between practice and theory. Right? So, the first difference is that we have gotten rid of qi, we are just going to go over the random variables one by one. Now ideally what would have we have done, late we were have taken, v1 sample done, new value for V when keeping everything constant. Right?

Then sampled a new value for v2 keeping everything constant, then sampled a new value for v3keeping everything constant, do we need to do this one by one, when I am sampling a new value for v2, do I need to wait for the step where I had sampled a new value for v1, the answer is no, why? So, I'm actually asking you the question v2, given all the other random variables. Okay? That means V 1, V 3 up to VM and H 1to HM, now I don't care what these values are. Right? Because anyways my VI is independent of each other. Right? So, I can just sample v2, given the configuration of the edges. So, even if myv1 had changed, in the first time step, it does not matter because this probability, is not going to change. How many if you get this? So, what does that mean instead of changing one visible variable at a time, I can, change all of them together. Right? So, I can do so, at every time step, instead of changing one of the visible variables, I can just change all of them together. So, the same as running, M steps of my Markov chain in one time step, I had M visible variables, I should have changed them one by one. Right? It would have taken me M steps, to go all the visible variables, but that's just a waste of time, because I don't need to, wait for one visible variable to change, to decide how the other visible variable is going to change, because the other visible variable does not depend on this visible variable, is that clear. Right? That's why I can do this block, Gibbs sampling, I can just change all the visible variables in one go. Right? So, that means that for every time step, I'm actually running M steps of the original Markov chain, because I have done these M transitions in one go. How many of you get this perfectly please raise your hands up and I? The ones who don't get this I see a very less population of that, but what's the problem. Okay? So, everyone gets it so this is what block Gibbs sampling does. So, remember that when I say one step, actually I have run the chain for M steps because I have done these M transitions. Right? At one go fine. So, what I am going to do is? First transition all the N hidden variables. Okay?

Refer slide time :(16:50)



Then transition all the M visible variables. So, I said the order was go all the visible variables and then go all the it didn't variables what I have written here is just the opposite of that but, it doesn't matter it the order could be anything Right? So, I just sample all the hidden variables, I sample all the visible variables and I'm going to do this for a large number of time steps that is a total of k plus r time steps. Okay? So, we have run the Gibbs chain, now what am I going to do? At the end of this I have some our samples, which I believe, have come from the distribution that I care about. Right? Now the first thing that I want to do is I want to compute this gradient, I want to do this update. Okay? Why is that a minus a plus instead of minus you are maximizing the lock like you're not minimizing it. Right? So, that's why it's a gradient descent and not gradient descent. Okay? Now I'm going to substitute, this. Okay? So, this plus sign expectation and now I've replaced the expectation, by this empirical explained expectation, based on the are samples that I have drawn, is that clear, how many you have fine with this step please raise your hands up and I. Okay? Good. And what's the next step going to be, be again it had that expectation, I'm going to replace that expectation, by its empirical estimation, again see had this expectation, I'm going to replace the expectation, by its empirical estimate is that clear. Okay?

So, this is what you will implement for training RBMs, this gives you the entire algorithm if we keep running this loop. So, what I have done is here, is what am I running here stochastic gradient, descent minimize gradient, descent batch gradient, descent stochastic batch mini batch, everyone. Right? I'm going over every example in the training data and the update is inside this loop. Right? So, for every training instance, I'm making this update. Okay? And you also need to be clear about something, in this entire computation, which is this, this is my key computation and that's the computation, which is updating the gradient, where is the training instance being used here, I was just drawing some samples from the Gibbs chain. Right? Where is the training instance being used here? Which is a training instance here, remember this first expectation was actually H given V, what was this given V? The training sample that was the given V. Right? So, whatever we use here, in the first term which I've called,' VD', is actually a training instance. So, the first expectation uses the training instances, the second expectation uses the

sample instances. Is that clear? Okay? Everyone gets this. Okay? So, that ends our discussion on training our beams using Gibbs sampling, now a very simple thing on training our beams using contrastive divergence. So, a question was that again I am anyways doing this n steps here. Right? But the point is this n steps can be done together. Right? So, that's what the block Gibbs sampling is about. Okay? Is this for clarity of notation that I have written it yes. So, that is that is also done, you can reuse from wherever you had stopped. Right? So, in practice that would be more meaningful, because that means, so, is a very valid question. Right? So, you are running this no, no the weight, there is a problem there, at every point you're drawing from this distribution, this distribution depends on the model parameters, the model parameters are changing, at every training instance. Right? So, the distribution is changing. So, you need to start it from fresh. Yeah, new mark of change, yeah. Right? Okay? Right.