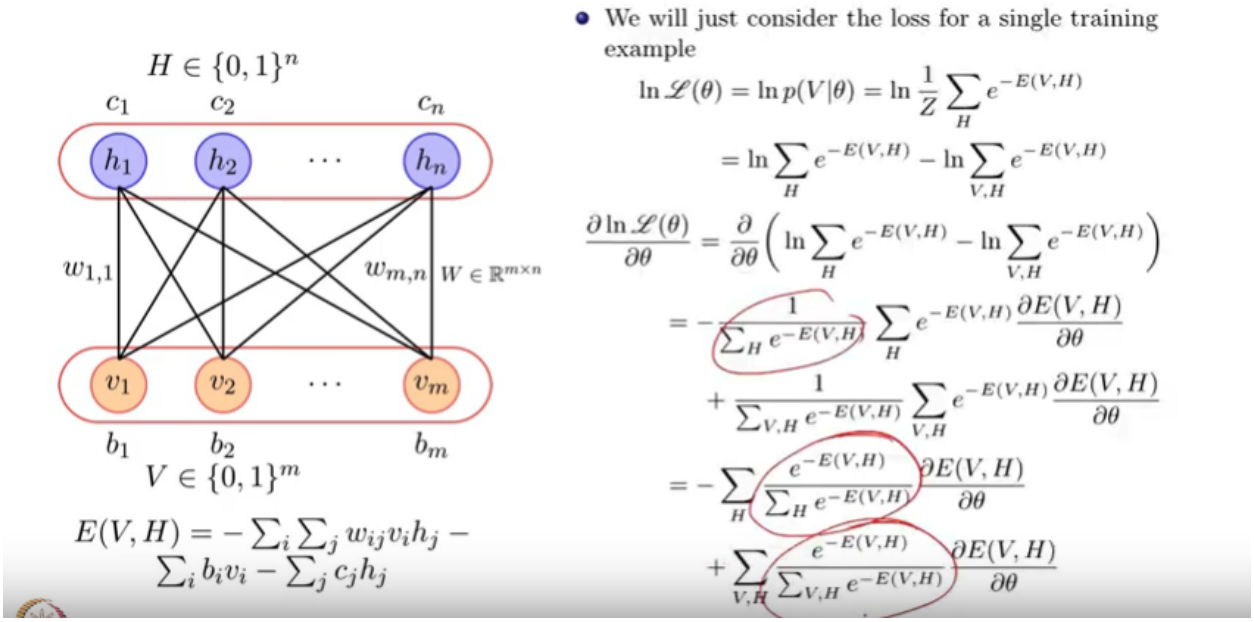
Lecture - 18.6

Deep Learning Part - II
Computing the gradient of the log likelihood

Refer slide time: (0:14)

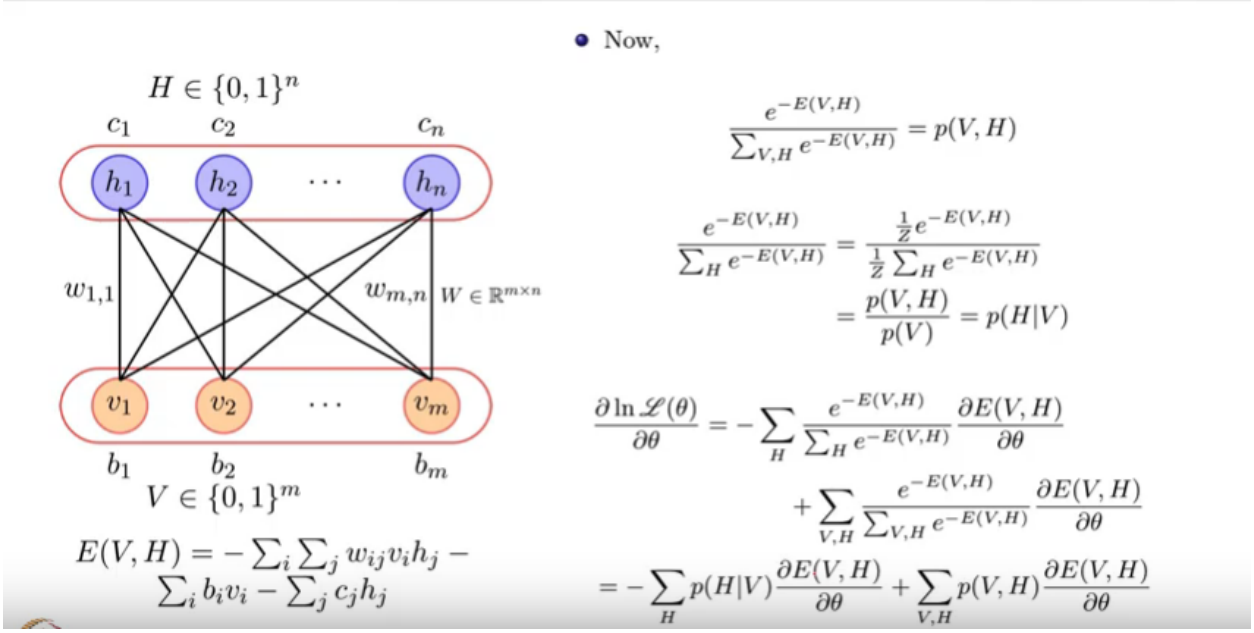


So, that's what we look at, we want to compute the gradient of the log-likelihood, with respect to the parameters of R model. So, let's just consider a loss, with respect to a single training data. Right? So, remember the loss is a summation over, all the training examples, I'm just going to consider it, for a given single training example. Okay? and this probability, I can write it as this, what am i have done here? what's that operation? marginalizes. I have marginalized over all the hidden variables. and what does Z? partition function just to make sure that. this visa probability distribution .So far so good. Okay? and actually Z is a summation over what? summation over V gamma H. right? it's summation over all possible values of V gamma H and here, you already have a summation of, all possible values of H. okay? so, now let's just kind of, tail this equation apart further and keep putting in more, details. So, this is log of a by B, so I've written it as, a log of a minus, log of B. where the B part, I am writing it as now, that summation over V gamma H. so, this is essentially Z, which is summing over all possible values of V gamma H. so how many terms are there in this summation? how many terms are therein the summation? M into n, all possible configurations of V Gamma H. how many terms are there in the summation? 2 power M plus N is that clear. how many terms are there in the first summation? how many terms are there in the first summation? 2 power N. Okay? fine.

So let's, start taking the derivative, with respect to theta. So, this is what it will look, so this is derivative of log of something, so what would we have, 1 over something, into derivative of that. Right? and thence have derivative of exponent of something, so what would that be, exponent of something, into derivative of that something. Right? that's how it's going to proceed. so can I directly write as this, you have won over something, which was this guy, then you have the exponent repeated, because they rate of e raise to X is, e raise to X. and then finally you have the, derivative of the energy function. and you have this negative fun, negative sign which I have taken, outside. Right? you have this min use of eh here, so I have taken that outside. and the same thing for the, second term also. I just want you to stare at it for 30 seconds and be sure that, you are comfortable with it, it just reorganize things, I've just taken this denominator inside. Okay? everyone comfortable with this, how many of you are fine with this, please

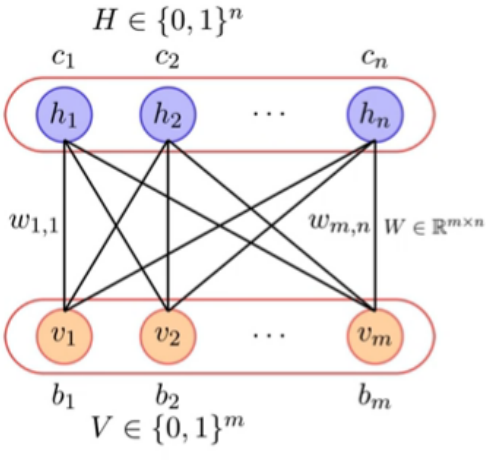
raise your hands, you guys are not fine with this, third row last three, time this okay. It’s too easy. Okay? now, let's focus on this and this, these are the two things that, I'm going to focus on that fine. So, first I'm taking the second guy. Right?

Refer slide time: (3:28)



That’s exactly what P of e Gamma H is; this is nothing but the partition function. So P of V Gamma H's 1 by Z into, the energy function right, exponent of the energy functions. How many forget this. Okay? And now, see having seen this, can you tell me, what this is going to be? I am asking you about, this quantity. The one in the circle P of V given H, I want everyone to answer, P of V given H. what’s the denominator? Okay? Let me make things easy for you. Right? I'll just multiply and divided this by 1 by Z. so, what's the numerator now? What’s the numerator? P of e Gamma H. what's the denominator? P of V. so, P of V Gamma H divided by P of V what is it? H okay? Fine. So, what we have is, that actually, the derivative takes the following form, I just I don't need to say, it again right. So, it says that it's P of H given V, into some quantity and then the other one is P of V Gamma H into some quantity. And this, atrocious summation outside it and both the cases, 1 over 2 raised to M plus n terms and the other over 2 raise to n terms Okay? Is that fine. Okay?

Refer slide time: (5:01)



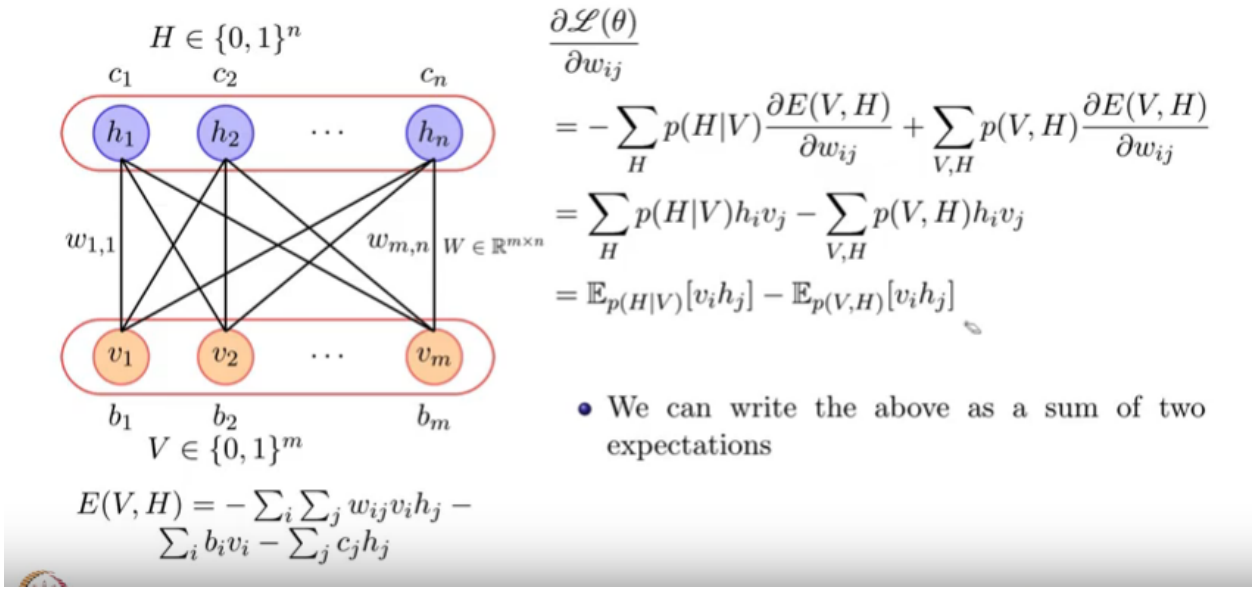
- Okay, so we have,

$$\frac{\partial \ln \mathscr{L}(\theta)}{\partial \theta} = -\sum_H p(H|V) \frac{\partial E(V,H)}{\partial \theta} + \sum_{V,H} p(V,H) \frac{\partial E(V,H)}{\partial \theta}$$

- Remember that θ is a collection of all the parameters in our model, i.e., $W_{ij}, b_i, c_j \forall i \in \{1,\ldots,m\}$ and $\forall j \in \{1,\ldots,n\}$
- We will follow our usual recipe of computing the partial derivative w.r.t. one weight w_{ij} and then generalize to the gradient w.r.t. the entire weight matrix W

We just speak into the next slide, I've just written this again. And Teta is a collection of all the variables that we have. Now, what is this quantity? Scalar, vector, matrix, tensor, teta is a collection of all the parameters that we have. Okay? Either call it a matrix or the vector I'd, if I mean, it depends on how you see, it if you just see all the parameters of the collection, as a vector then it's fine. If you think it's a matrix because, you have this kind of a neural network kind of a form, then it's also fine. Right? So, it's going to be some collection of gradients. And whenever you want to collect, a lot of partial derivatives, what do we do. What's our standard recipe? What do we do take? Take the derivative with respect to any one element and then kind of generalize toothed entire gradient. Right? So, we'll just focus on one of these weights, which is WIJ and from there, we'll try to generalize for the entire gradient. Okay? Fine.

Refer slide time: (5:51)

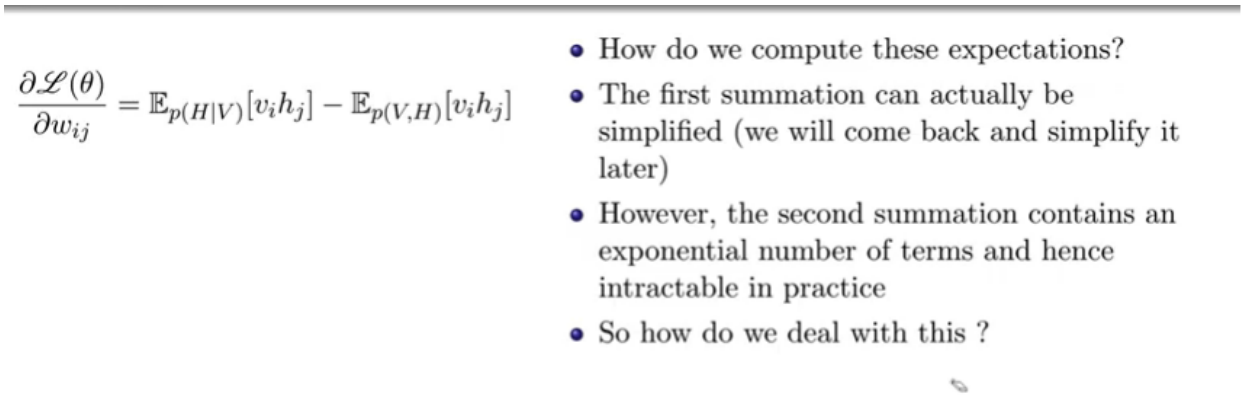


So now, I'm taking the derivative of the energy function, with respect to one. so just remember that, the D theta was completely inside the summation. so I'm now, just applying the replacing the D theta by D Wij, one of the weights. Everyone is fine so far right, no confusion at this point. Okay? Now, what's this derivative? V I minus V I H. right? so this - and that - would cancel, will I still have this summation, will I still have that summation, yes. that summation will let me right? This is the summation which is disappearing, right? you on the energy function, you have a summation, that's the one which is disappearing. And similarly for this guy, is this, it's the same. Right? So, you have it as this. Okay? You can actually write the above as, a sum of two expectations. What are those expectations? Mine and none, but what are those expectations? have to tell me these two things. What's the formula of expectation? Summation X PX. What am I asking you actually, what is X and what is px? so can you tell me, what goes inside the bracket? HIVJ and this is an expectation ,with respect to the distribution, H given V. and what about this? Same expectation with respect to V Gamma H. right? So, we can write this, as a summation of two expectations. So just imagine the bigger scenario. Right? that you have one training example, let's assume you are doing stochastic gradient descent, you have one training example, for one of the parameters, you have computed the gradient, of the loss function, with respect to that gara parameter, this is for one training example.

Now, to compute that gradient you need to compute this expectation, which actually sums over an exponential number of quantities, that means, for every stay step of stochastic gradient descent, you need to do an exponential number of computations. everyone realizes that, what is happening in the bigger picture ,you are running the loop for gradient descent and at one particular step of stochastic gradient descent, you need to do this gradient computation, in fact you need to do it for every step of gradient descent and at that point, you need to compute a summation, which requires an exponential number of computations, clearly this is not tractable. Right? This is going to be hard to do, but we have come up with, a proper formula, for what the gradient is supposed to be and now from, here on what will we do? What are, what are we always good at? Approximation. Right? So, we'll start doing some approximations, so that we don't have to do, these exponential number of computations. Okay? So is that clear, so now let's just focus on the overall paradigm that, we are sitting under. So, that we know exactly

where we are, so we go back to our data model parameters, learning algorithm and objective function. data what's the difference here, you only have the X's, you don't have device. Model we have chosen, RBMS as the model. Parameters are clear from the model itself, the objective function was maximize the probability of every instance that I have in my data. And the learning algorithm was back propagation with gradient descent. And now, to do gradient descent, I have this issue that I need to, compute these two expectations and it's not straightforward to compute them, because it requires an expected, exponential number of terms. Right? So, these are the two things. Now, this is the problem that, I need to solve, everything else fits in the story, but now I have to do gradient descent, how do I approximate this gradient by a few a number of computations. Okay? Fine.

Refer slide time: (9:53)



So how do we compute these expectations is the question? the first summation can actually be simplified and we'll come back to it, later on and we'll simplify it and you will see that, most of the terms there disappear, except for one or two terms, in fact one term. But, that still does not even if I simplify it, it still does not solve my problem, because the second term can actually not be simplified, it will still remain the same exponential number of terms. And the question is how do we deal with this? What do you do in such situations? No, but yeah, that's what we are doing stochastic gradient descent? But, for one step of stochastic gradient descent, we need to compute this expectation, which has an exponential number of terms. So, I want to avoid that exponential computed, computation. So what will I do? If you want a sample from the distribution? you sample from the distribution. Okay? Is sampling from this distribution easy, how many of you get that answer? You're saying that you need to sample from the distribution, what does that mean? How many of you understand what sampling from a distribution means? Okay? Quite a few, okay? So, the answer is correct. But let's look at a motivation for that. and also see why this is hard in the case of RBMS. Right? the answer is absolutely perfect, that you need to sample and approximate, the total summation by some samples, summation of some samples so that's a correct answer. why do we need to do that? And why is it hard for RBMS? These are the two things that I'm going to focus on today. And then we will solve the problem of; why it is hard for RBMS? Okay?