Indian Institute of Technology Kanpur

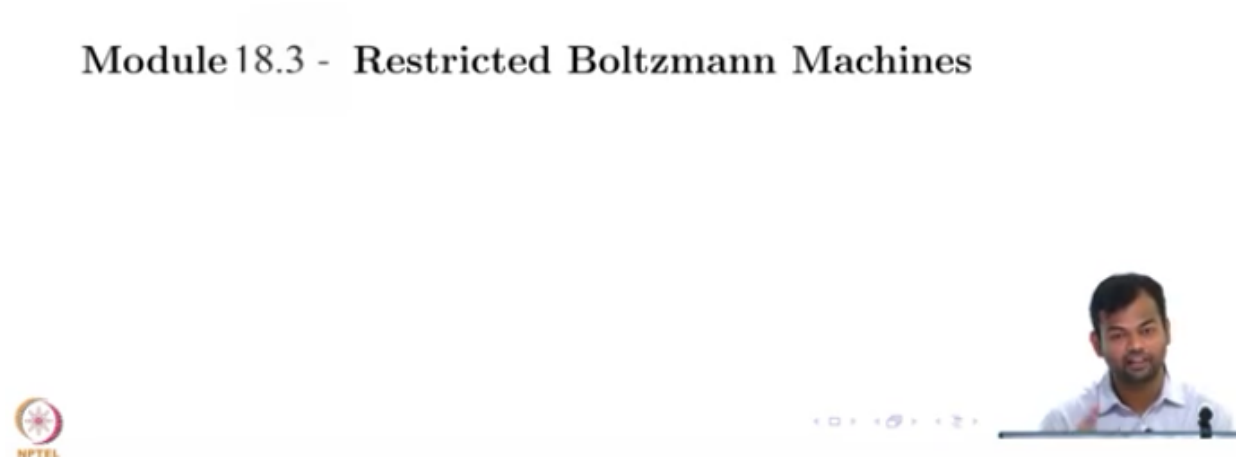
National Programme on Technology Enhanced Learning (NPTEL)

Deep Learning – Part II

Restricted Boltzmann Machines

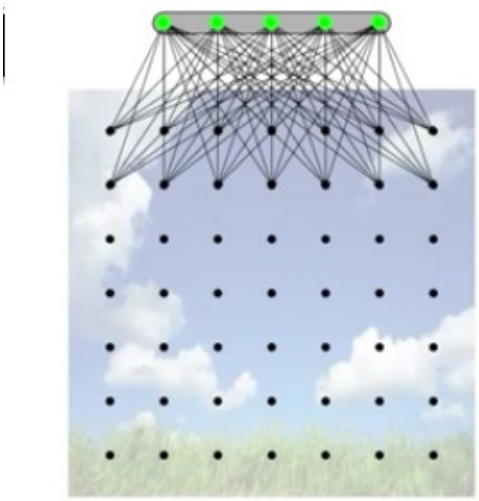
Module 18.3

(Refer Slide Time: 00:13)



So with that we are finally where we wanted to go which is restricted Boltzmann machines, I think we have all the background material ready now and we can now start discussing restricted Boltzmann machines.

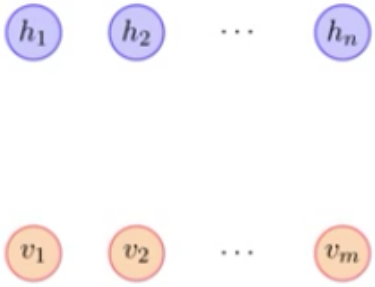
(Refer Slide Time: 00:24)



- We return back to our Markov Network containing hidden variables and visible variables

So we returned back to our Markov network containing hidden variables and visible variables, and everything that is written on the first bullet we understand Markov network, we understand what are hidden variables, we understand what are visible variables, and how they connect with each other to form this Markov network, okay.

(Refer Slide Time: 00:41)

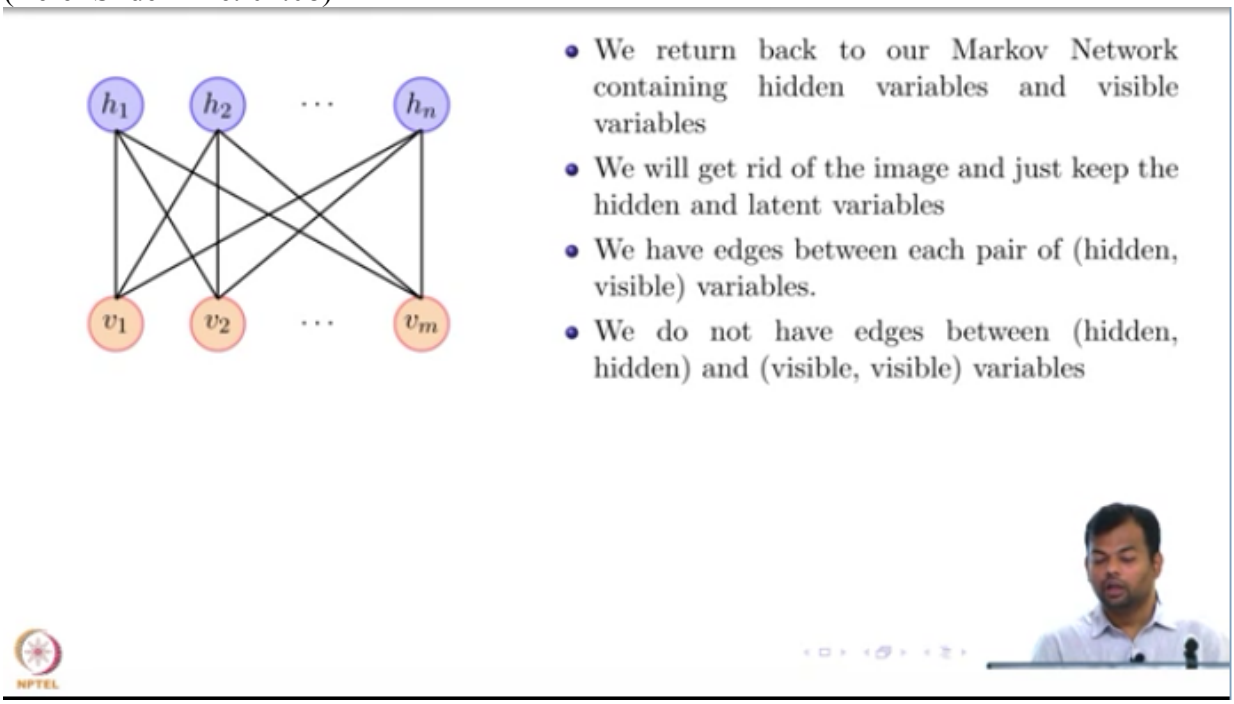


- We return back to our Markov Network containing hidden variables and visible variables
- We will get rid of the image and just keep the hidden and latent variables

I will get rid of the image, I will get rid of this image from the background, I'll just focus on the random variables that we care about which are all these N visible variables, and M hidden variables, right, so I will just, okay, we have changed this, this is N and M I think advantage

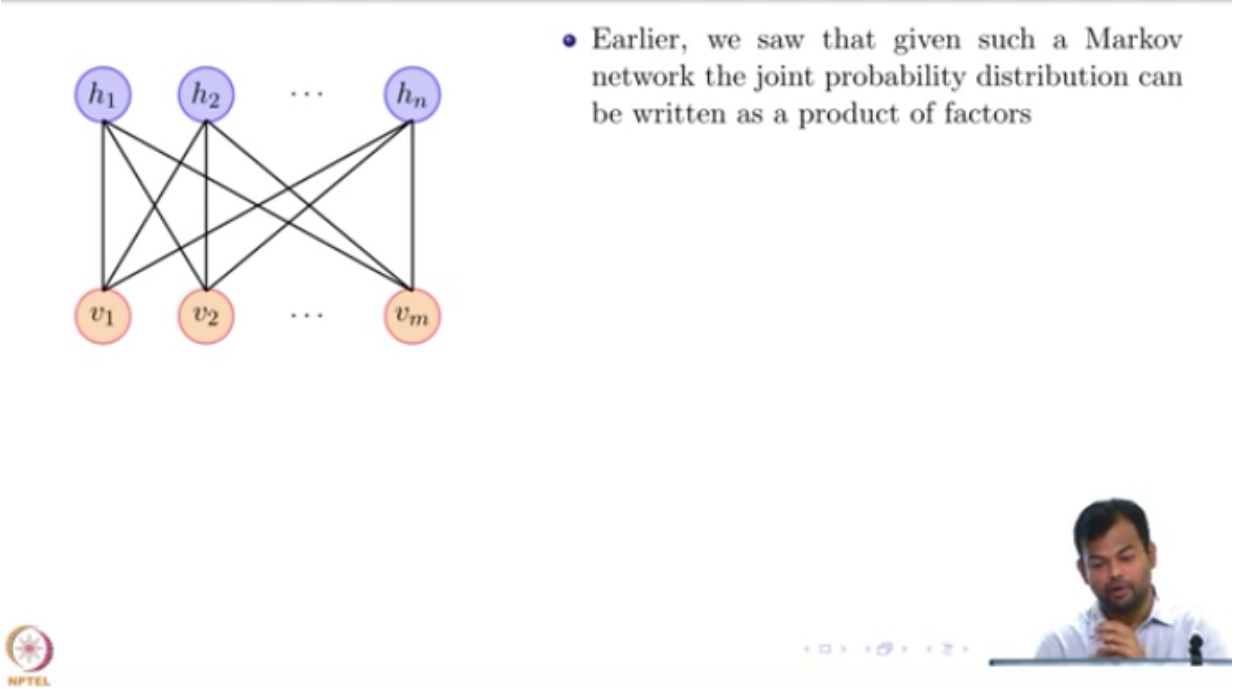
will have to change it, right, I hope this doesn't create confusion but there N of 1 and M of the other, okay.

(Refer Slide Time: 01:08)



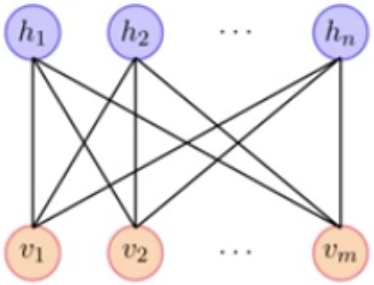
And as someone said this is a bipartite graph, so we have edges between each pair of hidden, visible variables, we do not have edges between hidden, hidden variables, and visible, visible variables, okay.

(Refer Slide Time: 01:20)

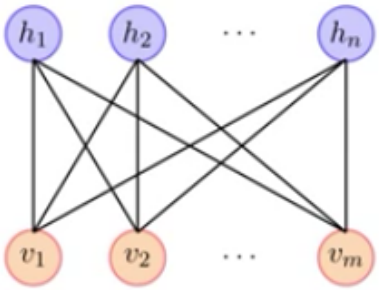


Earlier we saw that given such a Markov network, given such a means any in the Markov network you could actually get the joint probability distribution by writing it as a product of the factors. What are the factors in this particular Markov network? One factor corresponding to, what was the key term there? We have factors corresponding to, every one maximal clique, what are the maximal cliques here? Just edges, right, I mean just two nodes form a maximal cliques, so how many factors will you have? M x N, right, so correspond to maximal cliques, the maximal cliques in this case are just every pair of random hidden, this variables, so you will have M x N factors and this is how you can write it right.

(Refer Slide Time: 02:11)



- Earlier, we saw that given such a Markov network the joint probability distribution can be written as a product of factors
- Can you tell how many factors are there in this case?
- Recall that factors correspond to maximal cliques
- What are the maximal cliques in this case? every pair of visible and hidden node forms a clique
- How many such cliques do we have? $(m \times n)$

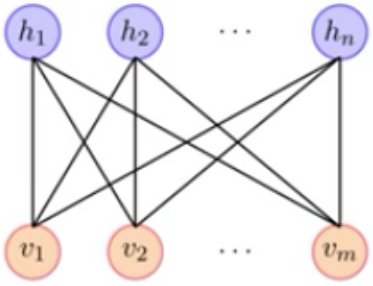


- So we can write the joint pdf as a product of the following factors

$$P(V,H) = \frac{1}{Z}\prod_i \prod_j \phi_{ij}(v_i, h_j)$$

For every visible variable, for every hidden variable you will have a factor which captures their interaction, okay, this is not, this particular factor is not a conditional probability distribution it just tells us these strength of the interaction between them, and then why do we have Z? Normalization so that we make it a probability distribution, okay. Okay, and let just remember that this belongs to 2 raise to M, this belongs to 2 raise to M, okay, fine, yeah.

(Refer Slide Time: 02:50)



- So we can write the joint pdf as a product of the following factors

$$P(V,H) = \frac{1}{Z}\prod_i \prod_j \phi_{ij}(v_i, h_j)$$

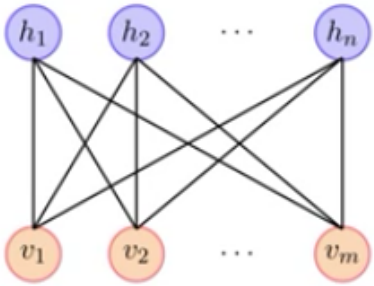
- In fact, we can also add additional factors corresponding to the nodes and write

$$P(V,H) = \frac{1}{Z}\prod_i \prod_j \phi_{ij}(v_i, h_j) \prod_i \psi_i(v_i) \prod_j \xi_j(h_j)$$

Now we can also add additional factors, at this point can anyone guess why I have added these two factors, what are these two factors? Corresponding to visible variables, and corresponding to hidden variables, right. Why I have added this factor? First of all is it legal to add these factors? You said that the factors have to be maximal clique s, it's desirable to have maximal cliques, it's not that we can have anything which is, we cannot have anything which is less than a maximal clique, right, so you got to think of each of these independent nodes and add factors for them.

Why would I be doing this? You have to go to the normal route, right, if you want to learn say VI, you will have to take P(V,H) and then marginalize over everything else and so on, right, so it doesn't mean that you just take this and say that this is P(V, I) right that's not correct, right, you see that, right, because it is specific reason why we have this and we will get to that.

So first of all let's understand this is legal to do this that as long as I ensure that my partition function is such that the whole thing gets normalized and I get a probability distribution I can add these factors, it's not wrong to do that, okay. I have factors corresponding to the maximal cliques and I have also decided to add additional factors to my graphical model, right, so that's perfectly fine to do that.
(Refer Slide Time: 04:13)



- So we can write the joint pdf as a product of the following factors

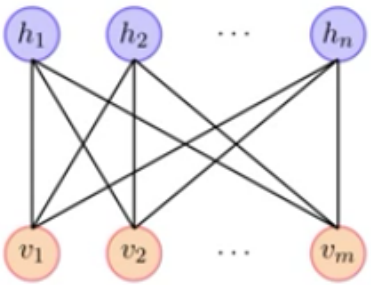
$$P(V,H) = \frac{1}{Z}\prod_i\prod_j \phi_{ij}(v_i,h_j)$$

- In fact, we can also add additional factors corresponding to the nodes and write

$$P(V,H) = \frac{1}{Z}\prod_i\prod_j \phi_{ij}(v_i,h_j)\prod_i \psi_i(v_i)\prod_j \xi_j(h_j)$$

- It is legal to do this (i.e., add factors for $\psi_i(v_i)\xi_j(h_j)$) as long as we ensure that Z is adjusted in a way that the resulting quantity is a probability distribution

And notice that this partition function is actually very complex, so what do I mean by summation V? How many elements are there in the summation? Is this a vectors, scalar, so that's very important, you understand what summation V means? How many terms is that summation have? It means sum over all possible V's, right, so this is actually summation, all V is belonging to capital V and how many such V's do we have? How many?
(Refer Slide Time: 04:53)



- So we can write the joint pdf as a product of the following factors

$$P(V,H) = \frac{1}{Z}\prod_i\prod_j \phi_{ij}(v_i,h_j)$$

- In fact, we can also add additional factors corresponding to the nodes and write

$$P(V,H) = \frac{1}{Z}\prod_i\prod_j \phi_{ij}(v_i,h_j)\prod_i \psi_i(v_i)\prod_j \xi_j(h_j)$$

- It is legal to do this (i.e., add factors for $\psi_i(v_i)\xi_j(h_j)$) as long as we ensure that Z is adjusted in a way that the resulting quantity is a probability distribution
- Z is the partition function and is given by

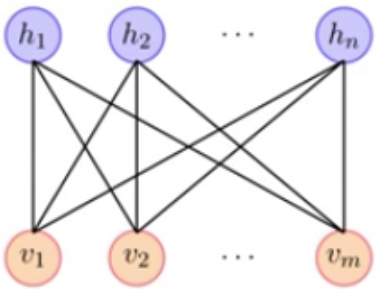
$$\sum_V\sum_H\prod_i\prod_j \phi_{ij}(v_i,h_j)\prod_i \psi_i(v_i)\prod_j \xi_j(h_j)$$

39/71

M, we have M such V's, I asked you what V is, you said V is the vector, and then you are telling me there are M such vectors.

How many of you say M? Now how many of you can give me the right answer? 2 power M, so I have to give you wrong answer to get the right answer, 2 power M right, this is all possible outcomes in your, so what does that denominator typically do in probability? It considers all possible outcomes, what are all possible outcomes? All possible values for the vector V, how many such values do we have? The entire space which is 2 raise to M, what about the second summation? 2 raise to M, right, yeah, so this is a summation over N exponential number of terms, so Z is a problem, right, this factorization all is fine to write, but still this is intractable because of this denominator which will have to go over an exponential number of terms, at some point will have to handle this that this denominator either gets out of the scene or it becomes attractable, one way or the other we'll have to handle this, right, is that fine, okay.

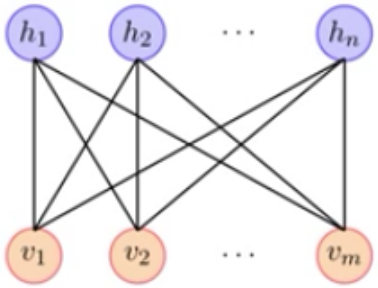
So that is understand each of these factors in more detail, for example if you have the factor phi 1 1, which essentially captures the interactions between V1, H1,
(Refer Slide Time: 06:13)



- Let us understand each of these factors in more detail
- For example, $\phi_{11}(v_1, h_1)$ is a factor which takes the values of $v_1 \in \{0, 1\}$ and $h_1 \in \{0, 1\}$ and returns a value indicating the affinity between these two variables

what will this factor actually give you? What will it take as input, and what will it give it to you as output? What will it take as input? The values of V1 and H1, how many such values are possible? 4, 2 x 2, for each of this 4 configurations it will give you one value, so what is this phi 1 1? Actually it's a table, right, very similar to the conditional probability distribution table where you have to have a value for all possible configurations of the two variables, right, so we could have a table like this and it could have some affinities in it, and that will give us some values, okay.

(Refer Slide Time: 06:53)

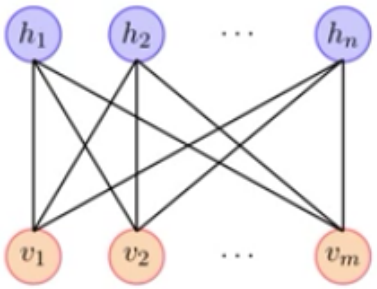


$\phi_{11}(v_1, h_1)$		
0	0	30
0	1	5
1	0	1
1	1	10

- Let us understand each of these factors in more detail
- For example, $\phi_{11}(v_1, h_1)$ is a factor which takes the values of $v_1 \in \{0, 1\}$ and $h_1 \in \{0, 1\}$ and returns a value indicating the affinity between these two variables
- The adjoining table shows one such possible instantiation of the ϕ_{11} function
- Similarly, $\psi_1(v_1)$ takes the value of $v_1 \in \{0, 1\}$ and gives us a number which roughly indicates the possibility of v_1 taking on the value 1 or 0

Similarly what would sai V1 be? We'll take one of these two possible values which is 0 or 1 and it will give you a output for that, right, so sai V1 could be this small, sorry, so we have this sai 1 and you could have a similar argument for the zeta as also right,

(Refer Slide Time: 07:05)



$\phi_{11}(v_1, h_1)$		
0	0	30
0	1	5
1	0	1
1	1	10

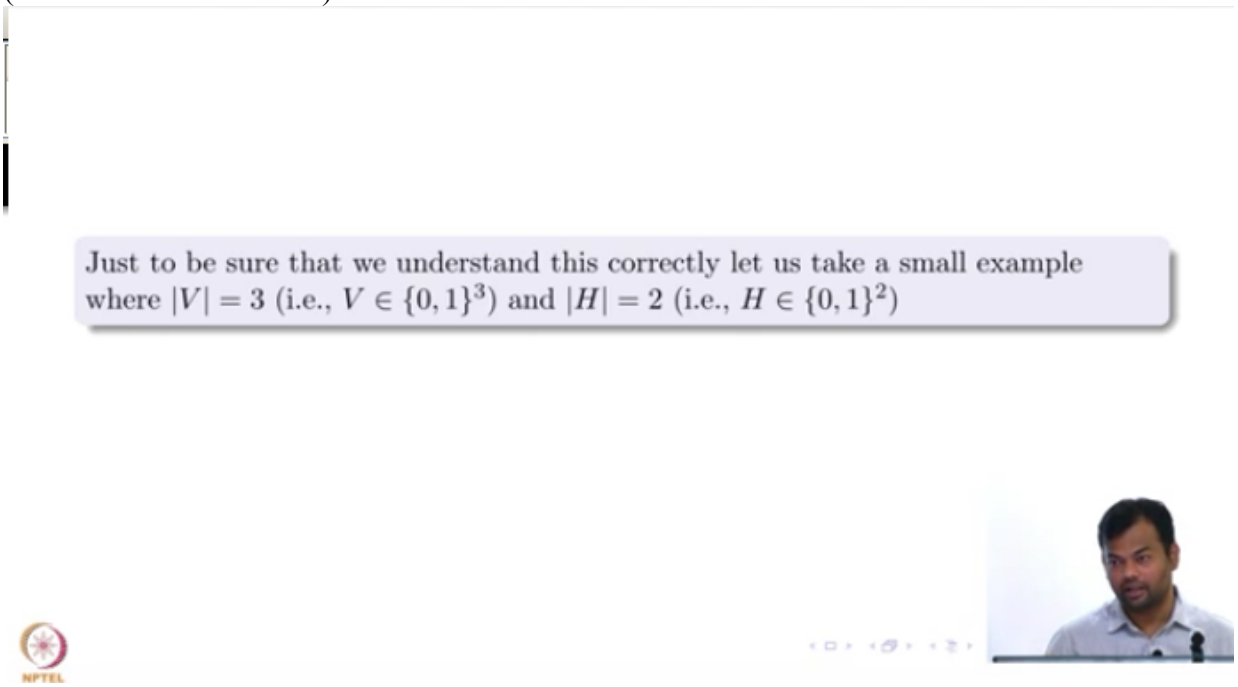
$\psi_1(v_1)$	
0	10
1	2

- Let us understand each of these factors in more detail
- For example, $\phi_{11}(v_1, h_1)$ is a factor which takes the values of $v_1 \in \{0, 1\}$ and $h_1 \in \{0, 1\}$ and returns a value indicating the affinity between these two variables
- The adjoining table shows one such possible instantiation of the ϕ_{11} function
- Similarly, $\psi_1(v_1)$ takes the value of $v_1 \in \{0, 1\}$ and gives us a number which roughly indicates the possibility of v_1 taking on the value 1 or 0
- The adjoining table shows one such possible instantiation of the ψ_{11} function
- A similar interpretation can be made for $\xi_1(h_1)$

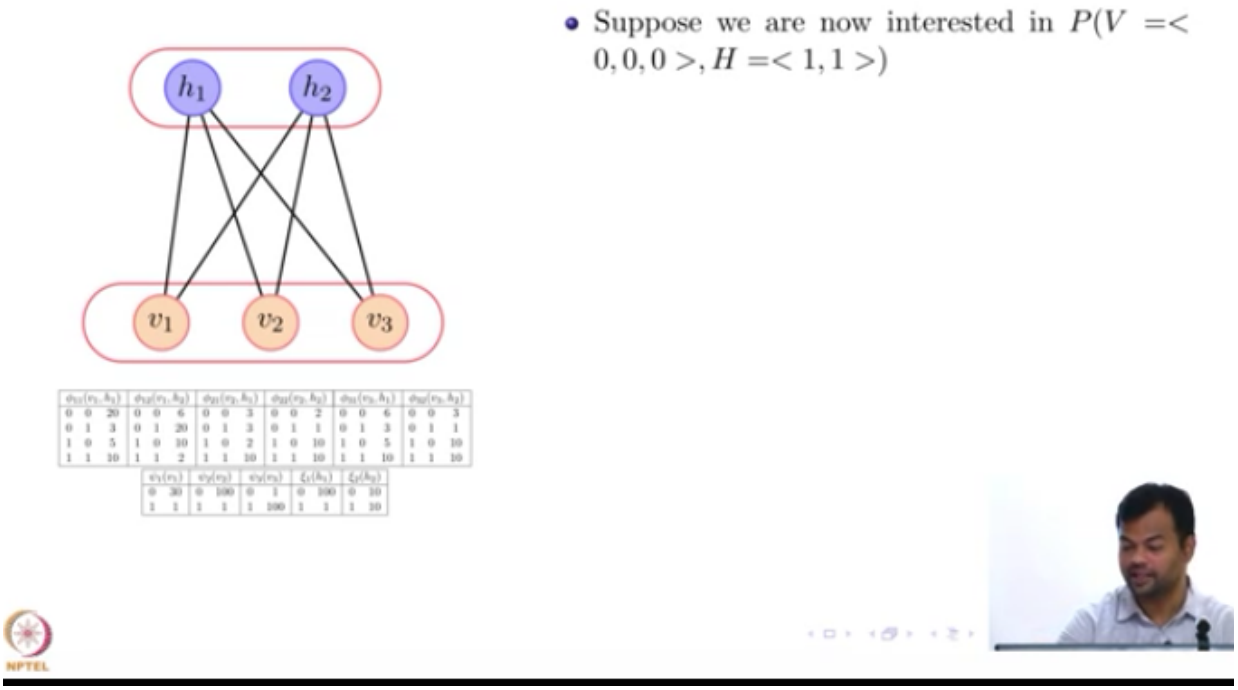
40/71

so you will have this zeta 1 also which looks something similar which takes on one of the two possible values for each random variable and engage it.

How many of these tables do you have in your joint distribution? Everyone, how many of these tables do you have? M, and how many of zetas? Now that I know its called zeta, okay. Okay, so just a bit short and we completely understand it,
(Refer Slide Time: 07:31)

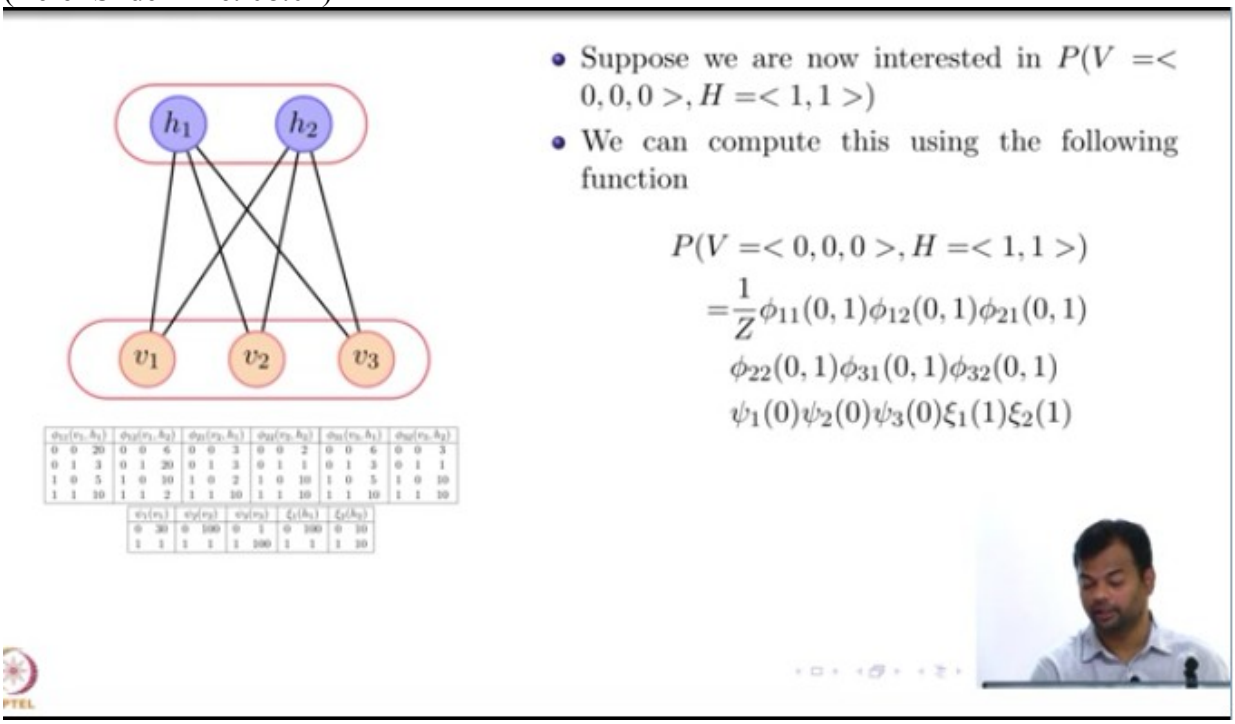


so let's take a very small example where we V = 3, and H = 2, and again both of them 0 to 1, R is to 3 and 0, 1, okay, so this is what we have.
(Refer Slide Time: 07:43)

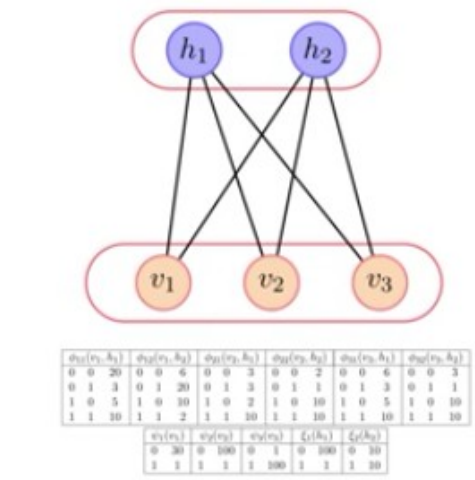


Now supposed we're interested in P (V=0, 0, 0 and H = 1, 1), how is this probability going to be computed? I have given you all the factors right, we always assume that there is some

miracle which gives us things, how are you going to compute this value? First you will write it down as a product right,
(Refer Slide Time: 08:04)



so these are all the factors in your product, to each of these factors you will supply the right configuration, right, so the first factor we will supply is 0, 1, second factor again 0, 1 and so on, and to the individual factors you will assign a pass 0, 0, 1, 1 and so on, right, whatever are the values that you have taken for these particular random variables, you will plug in those values in the factors, look up the table, get those values and then you have this hard job and now this should be clear that this goes over all possible values of all the random variables,
(Refer Slide Time: 08:37)



- Suppose we are now interested in $P(V =< 0,0,0 >, H =< 1,1 >)$
- We can compute this using the following function

$$P(V =< 0,0,0 >, H =< 1,1 >)$$
$$= \frac{1}{Z}\phi_{11}(0,1)\phi_{12}(0,1)\phi_{21}(0,1)$$
$$\phi_{22}(0,1)\phi_{31}(0,1)\phi_{32}(0,1)$$
$$\psi_1(0)\psi_2(0)\psi_3(0)\xi_1(1)\xi_2(1)$$

- and the partition function will be given by

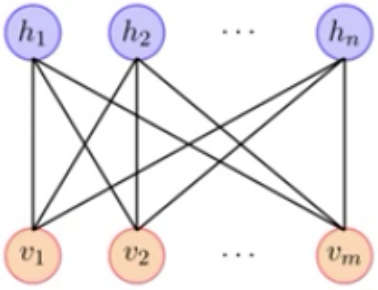
$$\sum_{v_1=0}^{1}\sum_{v_2=0}^{1}\sum_{v_3=0}^{1}\sum_{h_1=0}^{1}\sum_{h_2=1}^{1} P(V =< v_1, v_2, v_3 >, H =< h_1, h_2 >)$$

right, that's why this was exponential, right.

This is 2 raise to 3, this is 2 raise to 2, okay, so we'll compute the partition function also by plugging in all possible configurations and finally you will get the value.

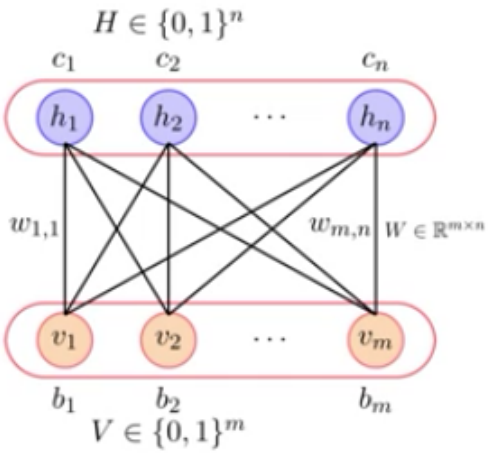
(Refer Slide Time: 08:52)

- How do we learn these clique potentials: $\phi_{ij}(v_i, h_j), \psi_i(v_i), \xi_j(h_j)$?



Now how do we learn these clique potentials? Finally getting to the question, whenever we want to learn something neural networks, some total of the course is 0, whenever you want to learn something what do we introduce? Parameters, right, so what we will have to introduce? We'll have to introduce a parametric form for these factors, okay.

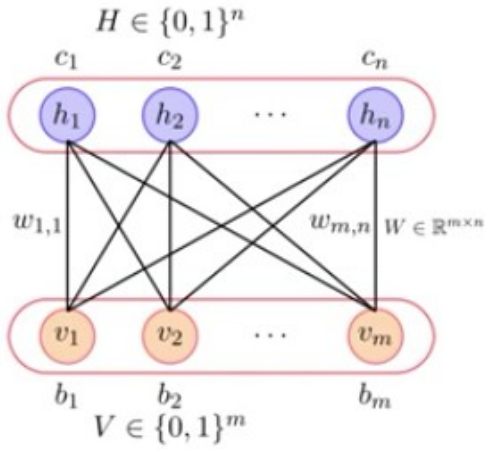
(Refer Slide Time: 09:22)



- How do we learn these clique potentials: $\phi_{ij}(v_i, h_j), \psi_i(v_i), \xi_j(h_j)$?
- Whenever we want to learn something what do we introduce? (parameters)
- So we will introduce a parametric form for these clique potentials and then learn these parameters

So we'll introduce a parametric form and specific parametric form that we are going to consider is the following.

(Refer Slide Time: 09:32)



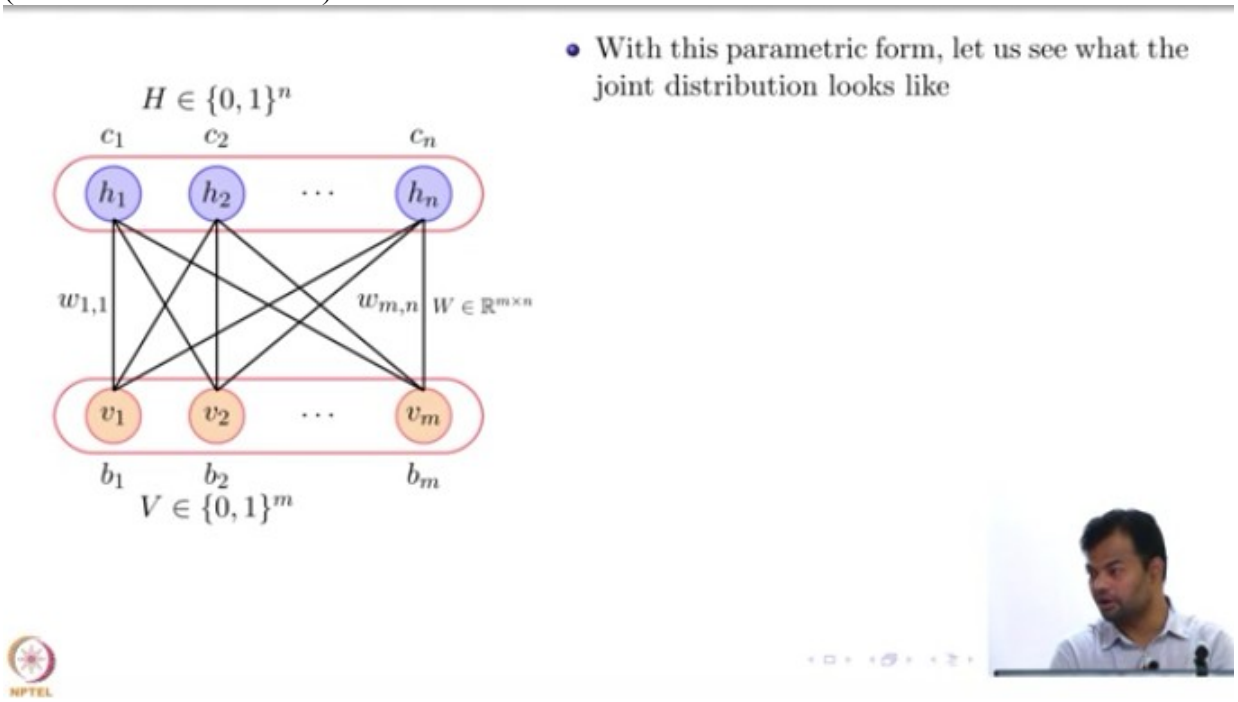
- How do we learn these clique potentials: $\phi_{ij}(v_i, h_j), \psi_i(v_i), \xi_j(h_j)$?
- Whenever we want to learn something what do we introduce? (parameters)
- So we will introduce a parametric form for these clique potentials and then learn these parameters
- The specific parametric form chosen by RBMs is

$$\phi_{ij}(v_i, h_j) = e^{w_{ij} v_i h_j}$$
$$\psi_i(v_i) = e^{b_i v_i}$$
$$\xi_j(h_j) = e^{c_j h_j}$$

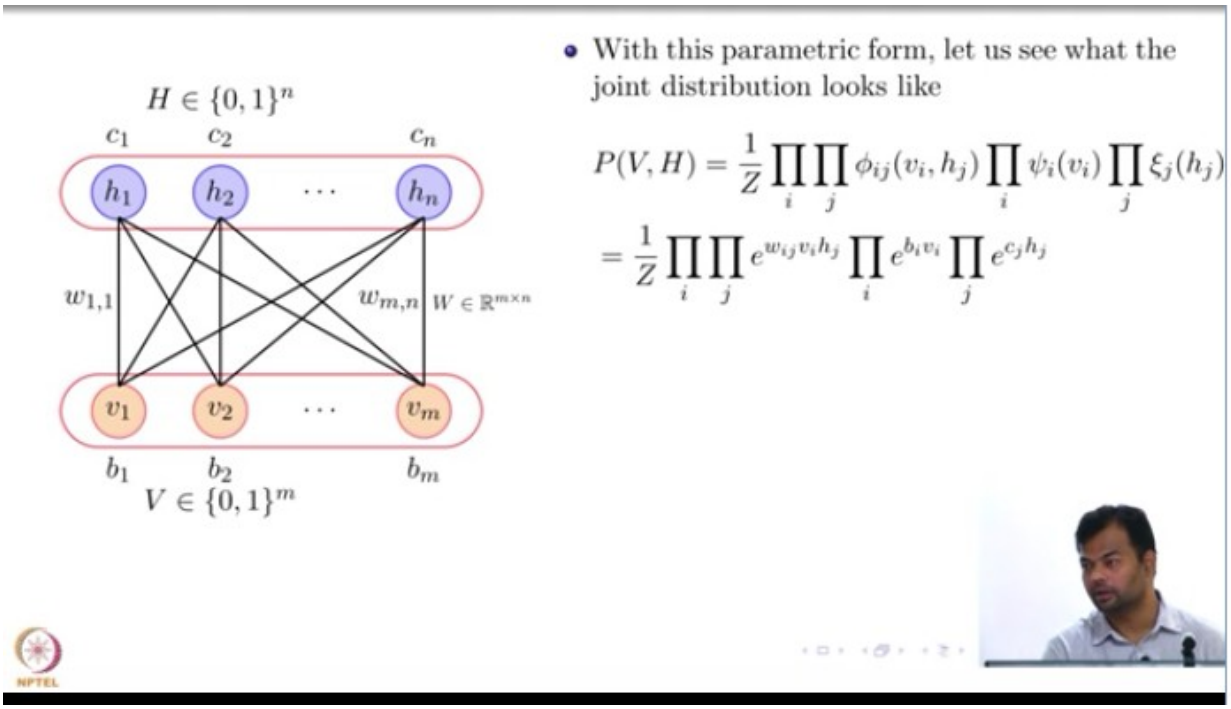
We are going to say that sai(VI, HJ) = E raise to, not this is not sai it's phi, okay, VI into HJ into some weight associated with that, so what's the parameter now? WHJ, so I'll learn this parameter in a way that some objective function is satisfied, I don't know what that objective function yet is, but I'll have to learn that so that is certain objective function is satisfied, and I've introduced parameters, okay.

Now these are BI's, VI's and I have introduced CJ and HJ, is that fine? This is again my modeling choice, this is how I have decided the factors should be, I could have chosen any other factors also, I could have said that actually it should be VI transpose, that doesn't make sense it's a single thing right, so I could have chosen some other parametric form also, right which is not necessarily a linear parametric form right, it did not be WIJ x VIJ, I could have chosen anything, this is the modeling assumption which is make that this is how my model looks like, this is what the parametric form is.

(Refer Slide Time: 10:47)

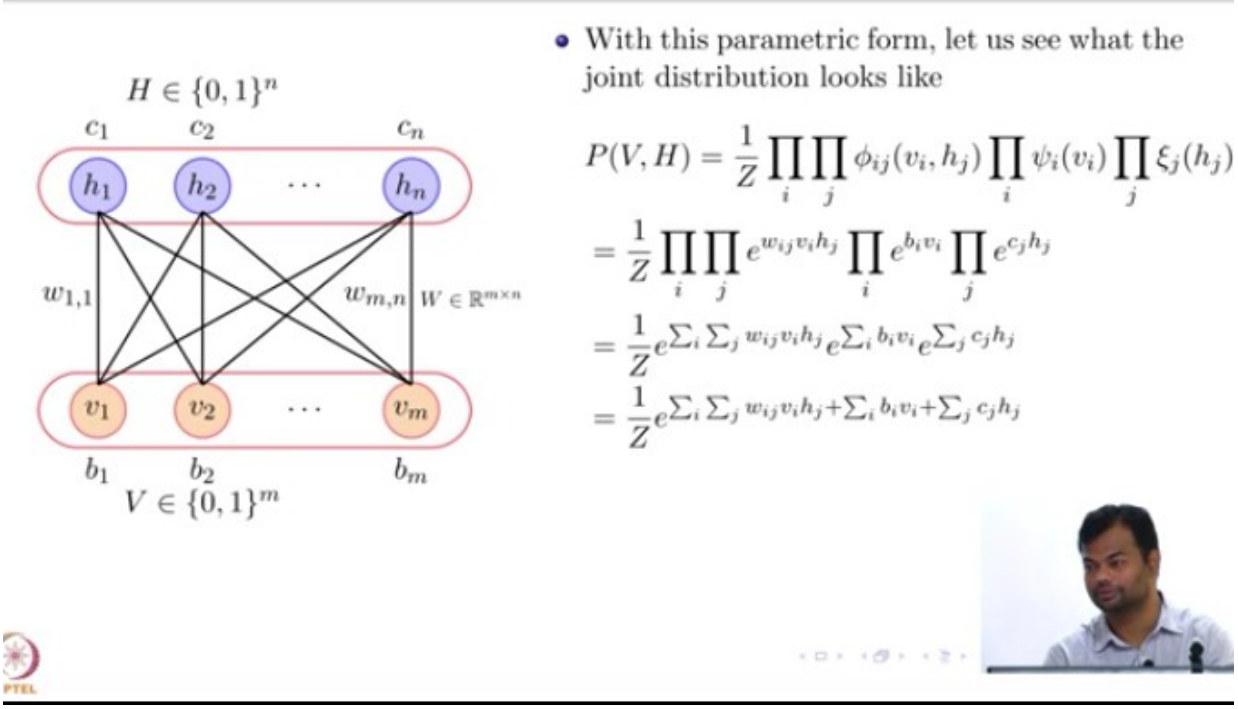


Now let's see what's the implication of this particular choice, so this is what our joint distribution looks like, now I'll substitute all these parameters into it what will I get?
(Refer Slide Time: 10:55)



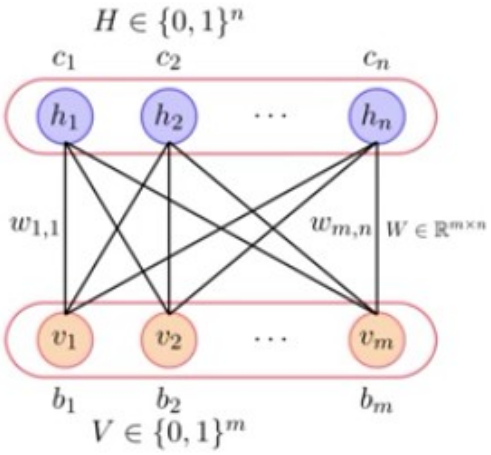
A product of many exponentials right, so this is what I'll get, okay, and a product of exponentials I can write as exponential of sums, is that fine? Have you ever seen this kind of equation before? Where?

(Refer Slide Time: 11:21)



Oh you've seen it in the assignment I thought you'll say it in the course, have you ever seen this in the course before? What is this? Looks very similar to, so it's very similar to the form WX + B + C, isn't it? Right, okay.

So now I can write, I'm going to write this as the following, it's E raise to minus of some function of V and H, it's right to write that because this is actually a function of V and H, right, all the V's and H is are contributing with that, right,
(Refer Slide Time: 12:09)



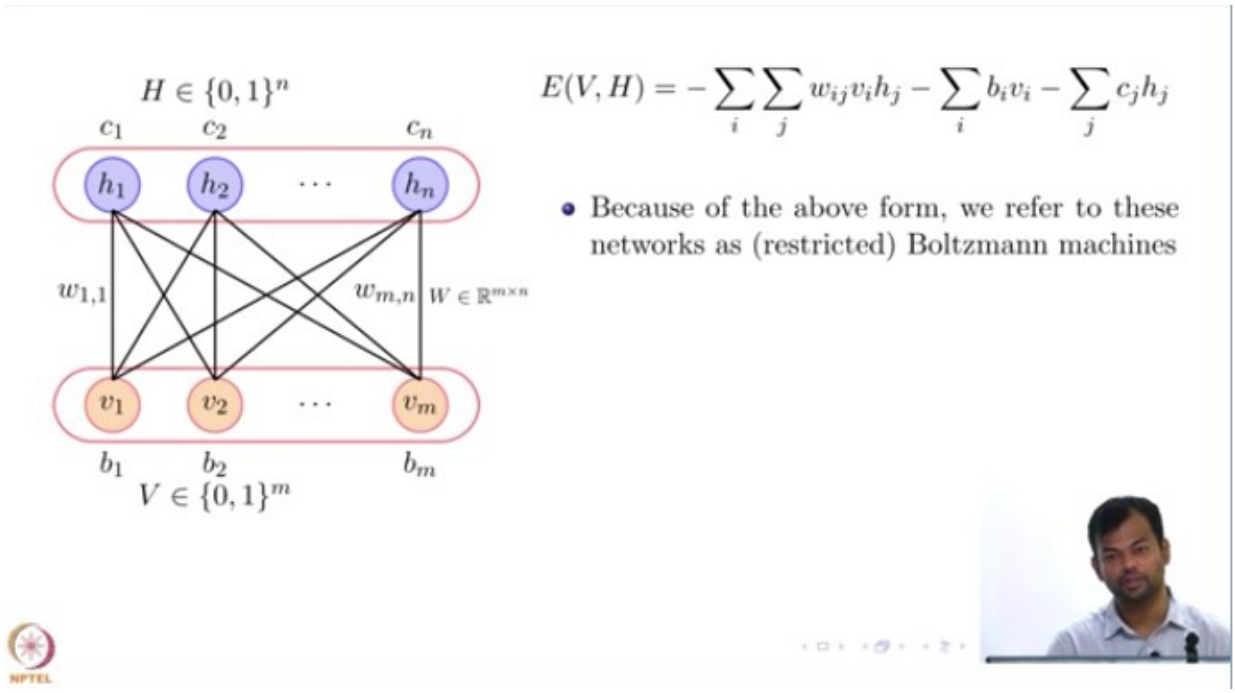
- With this parametric form, let us see what the joint distribution looks like

$$P(V,H) = \frac{1}{Z}\prod_i\prod_j \phi_{ij}(v_i,h_j)\prod_i \psi_i(v_i)\prod_j \xi_j(h_j)$$
$$= \frac{1}{Z}\prod_i\prod_j e^{w_{ij}v_ih_j}\prod_i e^{b_iv_i}\prod_j e^{c_jh_j}$$
$$= \frac{1}{Z}e^{\sum_i\sum_j w_{ij}v_ih_j}e^{\sum_i b_iv_i}e^{\sum_j c_jh_j}$$
$$= \frac{1}{Z}e^{\sum_i\sum_j w_{ij}v_ih_j+\sum_i b_iv_i+\sum_j c_jh_j}$$
$$= \frac{1}{Z}e^{-E(V,H)} \text{ where,}$$
$$E(V,H) = -\sum_i\sum_j w_{ij}v_ih_j - \sum_i b_iv_i - \sum_j c_jh_j$$

44/71

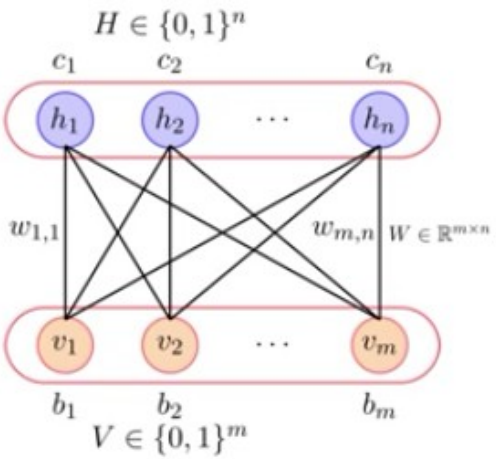
and where this function which is called the energy function is nothing but the negative of this sum that you see here, it's just some trickery to make the notation simple and make it consistent with something else in the literature, okay, so I have come up with a form which depends on V, H and the energy, this is known as the energy function, I'll tell you why the energy function is actually given by this and the probability distribution is a function of this energy function given by this, everyone clear with this? It is just trickery to arrange the terms and come up with this need form where the energy function is a linear function of some weights and your variables, is that fine, okay.

If the energy is high the probability would be low or high? Low, if the energy is the low, high because there is a negative right E raise to minus of energy, if the energy is high it will go into the denominator so it's okay, that's fine, okay. High energy low probability, low energy high probability, okay, have you seen this any time before?
(Refer Slide Time: 13:24)



Okay, so because of the above form we refer to these networks as restricted Boltzmann machines, why restricted? Why restricted? This everyone should answer, because we allow for only, remember the Naive Bayes model, now tell me why restricted? Because of all the possible connections you could had a complete graph right where every V is connected to every V, every H is connected to every H, you have restricted the number of connections, you have made it a bipartite graph and you're only allowed certain types of connections, hence restricted.

Why Boltzmann? Why Boltzmann? So this comes from actually statistical mechanics,
(Refer Slide Time: 14:17)



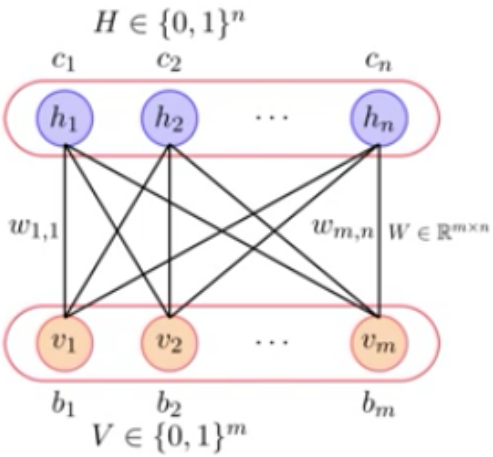
$$E(V,H) = -\sum_i \sum_j w_{ij} v_i h_j - \sum_i b_i v_i - \sum_j c_j h_j$$

- Because of the above form, we refer to these networks as (restricted) Boltzmann machines
- The term comes from statistical mechanics where the distribution of particles in a system over various possible states is given by

$$F(state) \propto e^{-\frac{E}{kt}}$$

whether distribution of particles in the system is given by this distribution, right, that a state is, the probability of a state is proportional to something of this form where E is the energy in that state, right, and we have a very similar form here, and that's why this is known as the Boltzmann distribution or the Gibbs distribution, okay, so we have, what we have done is we have taken our Markov network which was well motivated from the image example that the pixel actually comes from some Latin variables.

(Refer Slide Time: 14:45)



$$E(V,H) = -\sum_i \sum_j w_{ij} v_i h_j - \sum_i b_i v_i - \sum_j c_j h_j$$

- Because of the above form, we refer to these networks as (restricted) Boltzmann machines
- The term comes from statistical mechanics where the distribution of particles in a system over various possible states is given by

$$F(state) \propto e^{-\frac{E}{kt}}$$

which is called the Boltzmann distribution or the Gibbs distribution

And now we have made it possible to learn this distribution by introducing certain factors, the factors were very carefully chosen, which led to a certain form of the probability distribution which was of this form, E raise to minus of the energy, where the energy function was given by this which looks very similar to a neural network because it's WX+B, okay, so that's where we have finally reached restricted Boltzmann machines.

And now the question which I need to answer is, which is still not very clear, why are we doing this in the course on deep learning, right, what's the connection to neural networks, right, so we need to look at that, so we have restricted Boltzmann machines, we still are looking at it as a graphical model where you have certain sets of random variables and certain interactions and certain factors, but we still don't know why this is a neural networks, so that's what we look at next. Why are RBM's neural networks, okay.

Online Editing and Post Production

Karthik
Ravichandran
Mohanarangan
Sribalaji
Komathi
Vignesh
Mahesh Kumar

Web-Studio Team

Anitha
Bharathi
Catherine
Clifford
Deepthi
Dhivya
Divya
Gayathri
Gokulsekar
Halid
Hemavathy
Jagadeeshwaran
Jayanthi
Kamala
Lakshmipriya
Libin
Madhu
Maria Neeta
Mohana
Mohana Sundari

Muralikrishnan
Nivetha
Parkavi
Poonkuzhale
Poornika
Premkumar
Ragavi
Raja
Renuka
Saravanan
Sathya
Shirley
Sorna
Subhash
Suriyaprakash
Vinothini

Executive Producer

Kannan Krishnamurty

NPTEL Coordinator

Prof. Andrew Thangaraj
Prof. Prathap Haridoss

IIT Madras Production

Funded by
Department of Higher Education
Ministry of Human Resource Development
Government of India

www.nptel.ac.in