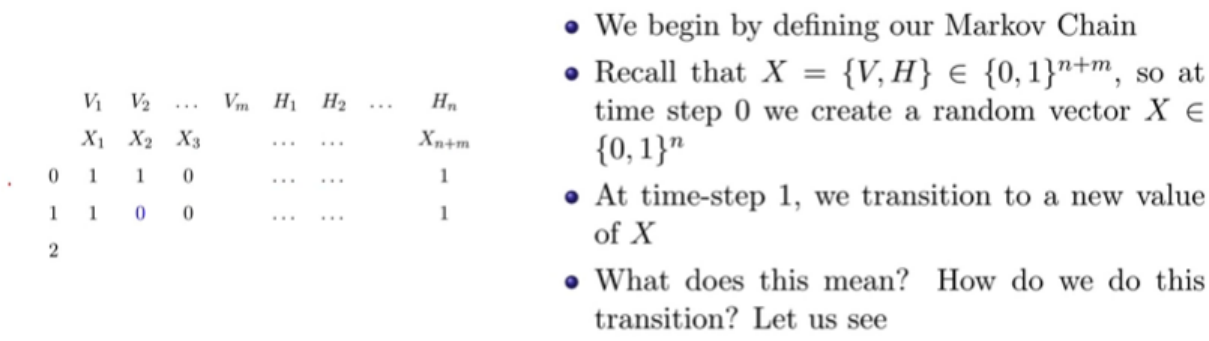
Lec 19.3
Setting up a Markov Chain for RBMs

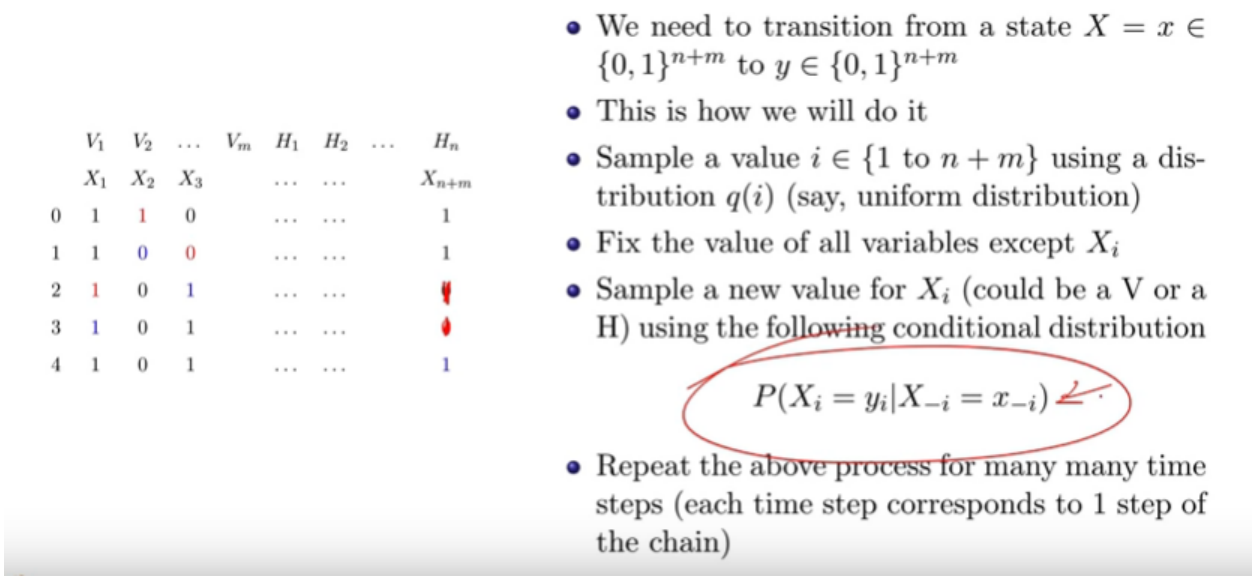
Let's start so, remember these questions the first one was, what the Markov chain is? So, we'll start with setting up the Markov chain, for RBMs. Disclaimer, for the next three slides, you'll not understand anything, why I am doing? What I am doing? But, it will become clear later on. Okay? So, let's just follow the procedure and then look then, then let's talk about, why we did that procedure? Okay?

Refer Slide Time :(0: 37)



So, we begin by defining our Markov chain. So, once again for clarity, our random variables, per v1 to VM and H 1 to Hn and I'm just renaming all of this as X 1 up to X n plus M. Okay? So, time step 0, I'll just initially as a chain with some, random variable, not random variable, sorry, some random value. Okay? So, I'll just sample some, value from the space that I have, 0 comma 1 raise to n plus 1 and say this is the vector which I have constructed. Okay? Now, I need to construct a Markov chain. So, I need to tell you what X 2 is going to be? What X3 is going to be? And so on. Okay? In fact I can tell you a Markov chain, completely if I tell you. What two things if I tell you? Initial distribution and transition matrix, for initial distribution I have just seemingly assumed a uniform distribution; I am just saying that I'll take any, Possible 'any value from the n plus M space. Okay? I've just taken the uniform distribution. Okay? And now, I have to tell you the transition matrix, which I'll not tell you. So, at time step one, we transition to a new value of x. So, this was the value of x, this was the value of x at time step 0. Now, at time step 1 we transition to a new value of x. Okay? So, what does this mean? What do I mean by transitioning to a new value of x and how do we do this transition? Wait so, let's see.

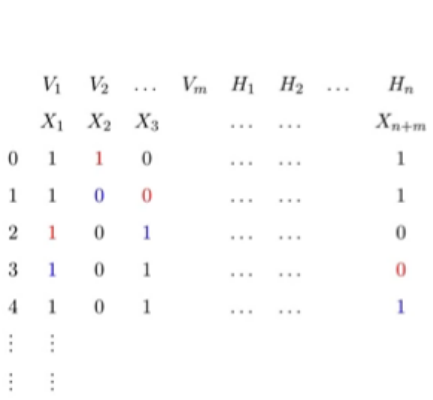
Refer Slide Time :(2: 01)



So, what do I need to do is? Given the value X equal to small X at time step 1, I need to transition to a new value, X equal to Y, at time step 2 that's what I mean from transitioning from one value to another. Right? So, remember that, the number of customers was transitioning from one particular vector value, to another particular vector value 8. So, that's how the transition happens? And the way I'm, going to do this transition, is the following, I'll sample a value, I ,from 1 to n plus M, where n plus 1 is the tote n plus M is the total number of random variables that I have, in my X. Right? My X, is a collection of random variables and there are n plus M of those. So, I'll sample a value of I, from this range and I'll use a uniform distribution. So, simple way of saying is I'll pick a number from 1 to n plus M. Okay? Now, I'll fix the value of all the variables, except X I. Okay? So, let's not avoid, let's avoid confusion here. So, here when I refer to X I, I don't mean the elements of the chain, I just mean the particular random variable, in X. Right? So, I mean, one of the constituents of capital X, is that fine, is there confusion with that No so, what I'm going to do is? Say I have sampled the value 2; I have picked up, 2 as the number here. So, I'm going to keep all the other random variables, fixed I am NOT going to change their value. Okay? And I'm going to sample a new variable, new value, for this random variable, using the following distribution. Okay? So, I'm going to sample a new value, for the picked, random variable using the following distribution. So, this is a fair question to ask. Right? I'm asking that, given the, values of all the other random variables, what is the probability of this random variable taking on a certain value, is that fine. So, does it make sense to sample according to this distribution? So, I'm telling you how I'm, going to define the transitions. Right? Instead of telling you the transition matrix. Right? Instead of giving you this to raise to n cross, 2raise to n matrix, I'm just telling a procedure for transition. Right? And of course eventually I will tell you how, this procedure actually gives you the matrix, I want to give you the matrix. But, I'm not giving it to you; I'm giving you a procedure. And now, I will show your one-to-one correspondence between this, procedure and the matrix but. Right? Now, we are just focusing on the procedure. So, the procedure is very, very simple, I have some configuration of the random variable X, at time step 1; I am going to fix all the components of this random variable, except one of them. Right? So, at time step 2, I'm going to keep all the values same, except for one of these. Now, whether I should keep the same for this, variable or change it, I will decide that, based on the following probability distribution.

I'm not sure whether I know this probability distribution or not, I don't know that yet, but, I am assuming that, this is some cloudy distribution according to which, I am going to decide, whether to change this value to zero or keep it one, because there are only two possible values Right? Again this is just a procedure; don't try to find too much meaning into it at this point. Okay? Now I'm going to repeat the above process, for many, many time steps, why I? Oh! Okay? X of minus I is not clear. So, X of minus I means, all the random variables, except the I at random variable. Right? So, this is, X I, given X 1, X 2, X 3 and not X I and all the other guys. Okay? And X minus I, this small guy, is a vector of dimension. The small X is a vector of dimension, every one, n plus M minus one; because I have dropped one random variable from there, we get fine, details you seem to have blanked out, is that fine. Okay? Sigh n plus M minus one. Okay? Is that here now. Okay? So, the notation I should have clarified. Now, repeat the above process for many, many time steps. So, what does it mean to repeat the process, again now, I am at time step one; again I am, going to pick up one of these random variables, see I picked up the random variable three, I'm, going to set, the value of all the other random variables, as it is, I am NOT going to change that and for this random variable, I am going to decide whether to change its value or not given the rest of the configuration, is it fine and I just keep repeating this and notice that, it's not necessary that the value will always toggle, it can actually remain the same also or not here, it can actually remain the same also. Right? So, at this time step again I picked one as the random variable which I am NOT going to, which I'm going to, sample again and fix all the remaining variables and at this time step based on this distribution, I actually ended up keeping the same value. So, now we have to live with this for the next 15 slides or so. Okay? So, yeah! The last column is constant it because we have not picked, any of those random variables, we are just working with these. So, the next infinite slides, just assume that the last column is all ones. Okay? Thanks. Clear it, you seem to be or we changed it at sometimes no, no, no we never picked this guy. Right? Unless we pick that guy it's not going to change. So, we picked this guy, is that clear, a picking one variable either changing it or keeping it the same, based on this distribution, is that fine, huh, ha, we will come to that, whether we are given this distribution not will come to that. Okay? So, that means you are asking the question that, is this computable or not. Right? And computable efficiently or not. Right? If so, then only I can transition. Okay? Right? So, again I just given you the procedure I need to explain the procedure, is it clear, this is how my Markov chain is, set up. Okay?

Refer Slide Time :(7: 59)

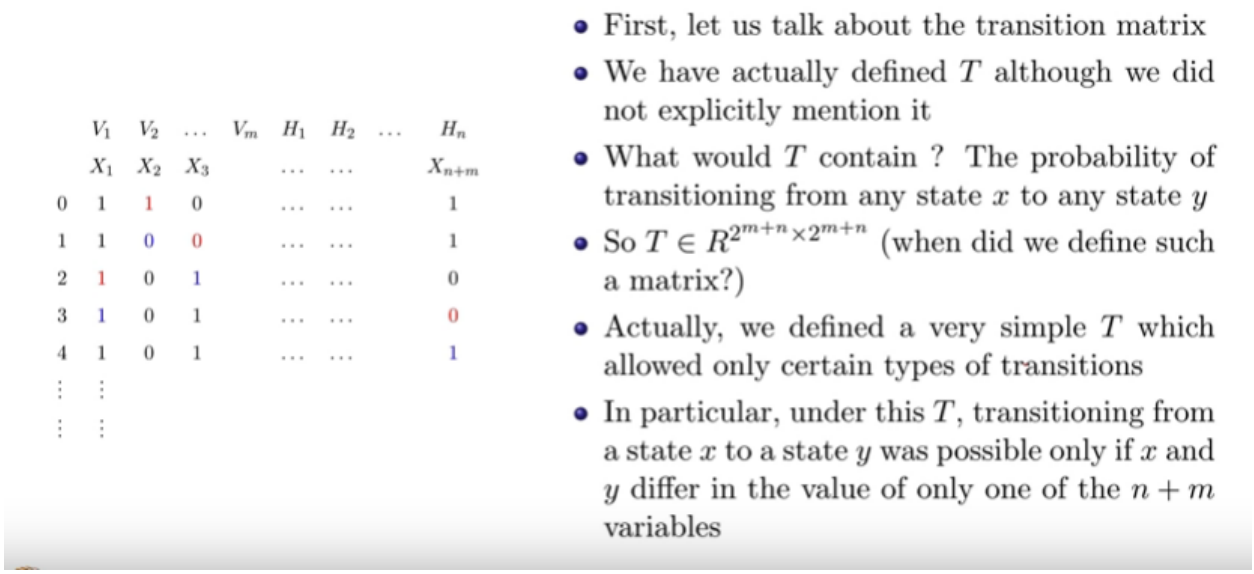


	V_1	V_2	...	V_m	H_1	H_2	...	H_n
	X_1	X_2	X_3		...	...		X_{n+m}
0	1	1	0		...	...		1
1	1	0	0		...	...		1
2	1	0	1		...	...		0
3	1	0	1		...	...		0
4	1	0	1		...	...		1
⋮	⋮							
⋮	⋮							

- What are we doing here? How is this related to our goals?
- More specifically, we have defined a Markov Chain, but where is our Transition Matrix T?
- How is it easy to create this chain (or creating samples $x_0, x_1, ...x_N$) ?
- How do we show that the stationary distribution is $P(X)$ (where $X = V, H$) [We haven't even defined T, then how can we talk about the stationary distribution for T] ?

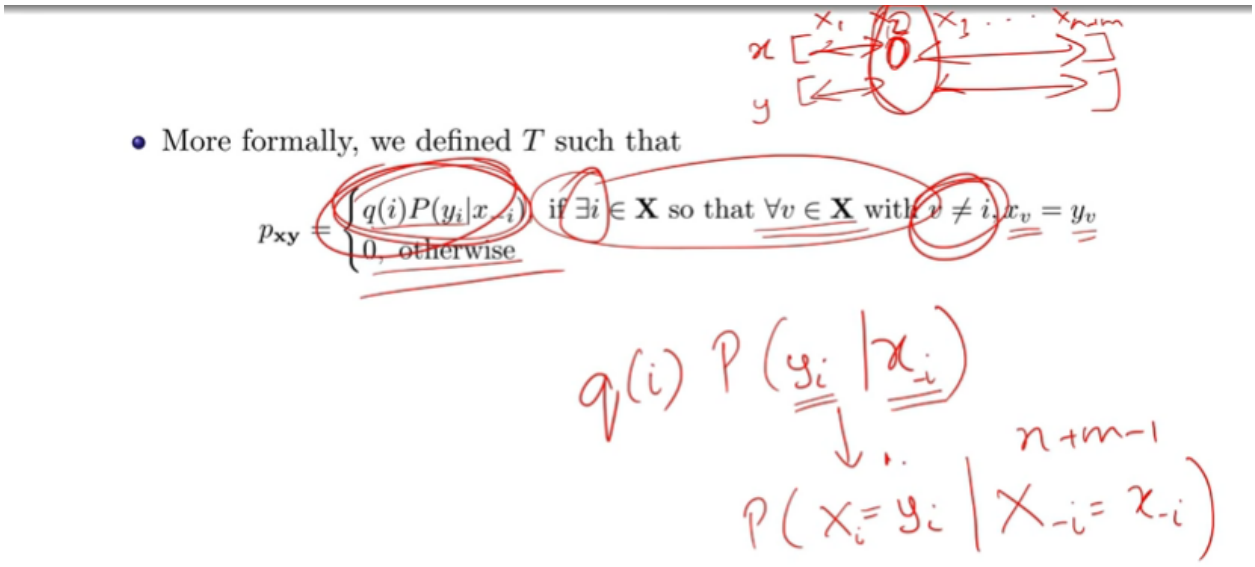
It's so, what are we doing here? How is all this related to our goals? Nothing is clear at this point. Right? So, but, we'll try to make it. Okay? Now we are violating the basic principle to define a Markov chain, I need to give you the transition matrix and I'm just shying away from that, I'm not giving the transition matrix, because I cannot give you one, because I cannot give you to raise to n plus M cause to raise to n person values. Right? Not in the duration of this course for sure. So, that's why I am NOT giving the transition matrix. But, we'll have to get to the transition matrix actually. How is it, easy to create this chain and that's the question which she was asking it, at every transition point I need to compute some probability, I need to compute that probability, I'm not even told you what that probability is, I need to tell you that, this transitioning is easy, I have not done that. And of course I have not even shown you, that the stationary distribution of this Markov chain is the distribution that we care about. Right? I've not even told you what's the transition matrix, is then the question of the stationary distribution does not even arise it. So, these are the three questions that I need to answer, I need to tell you what is the transition matrix explicitly in this procedure, I need to show that, it's easy to transition that means it's easy to draw these samples x0, x1, x2 across these time steps and finally I need to show you that, this indeed, has the stationary distribution, as a distribution that we care about, which is the Joint Distribution of our random variables. Right? Where X is equal to a collection of the visible, as well as the hidden believers. Right? So, you agree if I answer these three questions, then you'll not give me these blank faces that you're giving me. Right? Now, can you have a deal, yes or no everyone? Okay? Not everyone yet, but Okay?

Refer Slide Time :(9: 37)



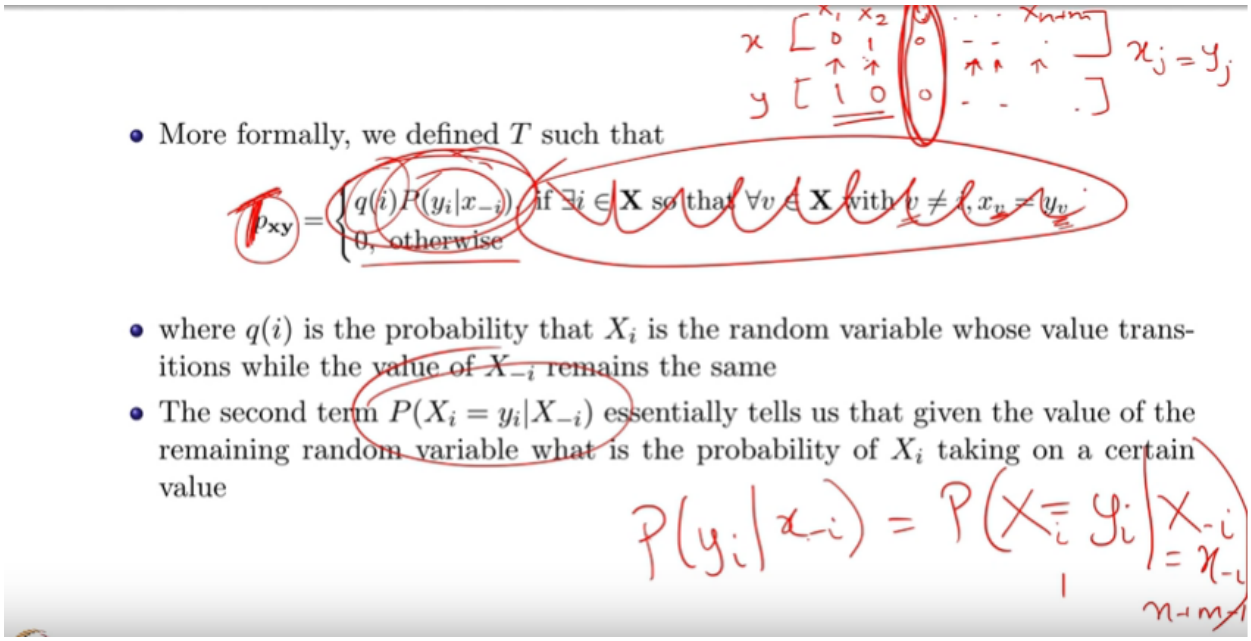
So, let us answer these questions one by one it's a first, let us talk about the transition matrix, we have actually defined at T, when I was giving you this procedure I've actually defined T although I did not tell it to you explicitly. And now, let me tell you what this T is, what would T actually contain? Like what would it actually contain? Quality of transitioning from state, a to state B for all possible a and B but that's what, it will contain it from any state X, to any state Y in my samples pace and my sample space is to raise ton plus M. Okay? At least we know what T contains? So, T is this kind of a matrix, when did we define such a matrix, did we define such a matrix? Did we define such a matrix? Actually, we did we defined a very simple matrix T and it allows, only certain types of transitions. Right? So, let's see, what it meets it, in particular under this matrix T, we allow only those state transitions, we differ from each other, at max then one random variable. So, that's the same as saying, T of XY, is equal to 0, if, do you get the question. So, we define the procedure in such a way that we allow, only certain types of transitions that's the same as saying that, I will not allow you to transition from state X to, state Y, if they differ in more than one random variable that means, T of X Y is equal to0, if x and y differ in their values for well more than one random variable, is it clear. So, you see I have defined a very, very sparse transition matrix and that's why I don't need to give you it, explicitly I don't need to give you the to raise to n plus m cross n plus M values, I have told you how what are the zeros and this. So, I just need to focus on the non zeros. And what were the non zeros? Did I give you a formula for then on zeros, did I. Okay? So, let's see.

Refer Slide Time :(11: 52)



So, we define T using this, I urge you to read it, understand it and explain it, to me focus on the zero otherwise that's always easy. So, let's see. Okay? So, let's see this is our capital X, adding this some problem here but, let's not tell them, say X 1, X 2, X 3 up to X n plus M. Right? So, it's saying that, P of X Y is equal to 0 and I told you when it is going to be 0. Right? So, the otherwise case is clear, for the first condition this is just a more formal way of saying that, there exists an I, well exist an umber I, belonging to 1 to n plus 1 M, such that, for all the random variables belonging to X, if the index of that random variable is not equal to high, then the value is the same across these two states, do you get that, how many if you are able to read this now? So, I have these two states. Right? I have the state X and I have the state Y, what this, cryptic looking first condition is telling me. Right? If there is exist a value I, such that, for all the things which are not equal to I, those state values are the same, for x and y. So, x and y remember our n plus m dimensional vectors. So, what it is telling is, n plus M minus 1 values are same and one of these values may or may not differ. Okay? If that is the case. So, that is how x and y look, then the probability of transitioning from X to Y, is given by this. So, now let's break, is that clear if the if condition ok, everyone gets safe condition please raise your hands, if you don't. So, you have a state X and state Y Okay? And so, what the first condition is telling is that? If there is one random variable, one index, for which apart from that index, all the other state values are the same. Okay? Now, why this particular form, what is this? So, what was Q I, Q I was the probability of picking that value I. Right? So, that has to be so, I have picked the value I and then, what does this mean? Actually, we I what is this actually a shorthand for, can you write the full formula for this, what is this a shorthand for? Are these random variables, these are values assignments. So, what is the full way of writing this? X equal to Y I, given or rather X I. Right? One of these guys taking on the value of I, I given that, all the remaining guys have a certain value. So, this is n plus M minus1 and this is of course 1, you get that. So, is it clear now, how I define the transition matrix, how many forget it that the procedure explicitly defined a transition matrix or rather implicitly defined a transition matrix? Now, please raise your hands, Okay? Fine good.

Refer Slide Time :(15: 16)



So, Q I was a probability of picking, I as that random variable, which is going to differ and all of the others are not going to differ. And P x I equal to y I, given the remaining random variables, I essentially tells us, what is the probability that the higher random variable will take on a certain value, given the values of all the other random variables. Is this easy to compute, what are our X's? What are our X's? X is actually equal to, X is actually equal to, everyone, it's a collection of V comma H, this is some conditional distribution defined on V comma H, you get that, do you think it would be easy to compute that, not sure so, we will get to that. Okay? But, at least are these definitions clear to you, what each of these quantities that you see mean, this if condition is clear to everyone, please the other hands, if you don't understand that, you not understand anything for the rest of the course. No, how many if you don't understand? Please don't raise, so many answer, I mean you do if you have if you do not understand, but, how many few don't understand? Fix it, you don't understand? What were you waiting for? Okay? What do you not understand? No, no that guy, you have bigger problems actually huh, the both the circles you don't understand. Okay? So, what I was interested in this P of X Y. Right? Okay? And it would have been better if I'd call this T of X Y. So, I am interested in the probability of transitioning from a state X, to a state Y. Okay? So, let's have this as X and this vector of course has the assignments for all the random variables that I care about, up to n plus n, that's some assignments. Right? So, these are zeros ones, zero something, something, something. Now, I'm interested in transitioning to a new state Y. Okay? Now, what the else part means zero otherwise tells me is that, if x and y, are going to differ in more than one random variable. So, I, I have Y such that, it's actually try to erase it with my finger. Okay? If I have a Y which looks like this. Right? So, it's different from X in two values, such transitions I was not allowing, if you remember the procedure, I was only allowing you to change one value, that's the same as saying that if x and y differ in more than one value, then the transition probability is zero, is that clear that's what the else part says and if you understand the else part? Part is just con; I mean the counter of that rate. So, what the if part is saying, this cryptic looking formula is saying is that, if there exists an I, belonging to n plus 1 m. Right? 1 to n plus 1 you have sampled 1 I, such that except for this I, the one which I have circled, for any other v,

any other index which is not equal to I, X of V is equal to Y of V. Right? So, these are all the indices, the ones which I am now, highlighting with an arrow, are all the indices which are not equal to I, for all those indices, X of J, let me call them, 'J' is equal to Y of J, is it clear, that's the same as saying that, these two states, do not differ in these values, in this n plus M minus 1 values, they're only allowed to differ in this one value. Now, for this one value, whether to keep it same or different that's, defined by this prod distribution. Okay? So, first thing that I need to take about is, sample value of I, that's Q I, once I have done that and now, tell me what's the probability of transitioning to a new value, given all the other random variables have been fixed. So, that's what this condition tells me. Right? And what I was telling you is that, that's actually a short form. Right? Is that clear I. Okay? Where X minus I is all the other, random variables in your collection, how many random variables do we have? N plus M. Right? So, this condition is on n plus M minus 1 values and what you have here is, one random video. Okay? Fine Okay? So, is that clear to you at this point? Okay? Now, Q I, my claim is its straightforward, because we assume it's a uniform distribution. So, drawing a sample from this distribution is not hard, what I need to show you that this other quantity is also, easy to compute. Okay? And we'll get that, that's what I'm going to show you know. So yeah! This is messed up, just focus on the explanation that I gave. Okay? I did not see it properly; I should have done that. Okay? Just focus on the expression which I gave you, if that is fine then I will just change this condition to match that. Okay? Yeah! but, all this dries, right from starting the description of RBMs, I had said that we are only going to focus on binary variables, there is, we will not be covering that in this course, we will have to assume certain forms for this distribution and so on it. But, right Now, we are not interested. Okay? But, I've still not shown you that this is easy by the way that I need to, to show you that, P of X to X, yeah! That's the same formula. Right? So, in that case Y is equal to X I. Right? Yeah! Okay? Let's not confuse the others. Okay? Yeah I get your point. So, I think there should be a summation but, I will just get back. Okay? Anyone else? Okay? Is it clear; despite messing up the most crucial side of the lecture is it clear. Okay? So, what I have done now, I have given you the,

Refer Slide Time :(21: 26)

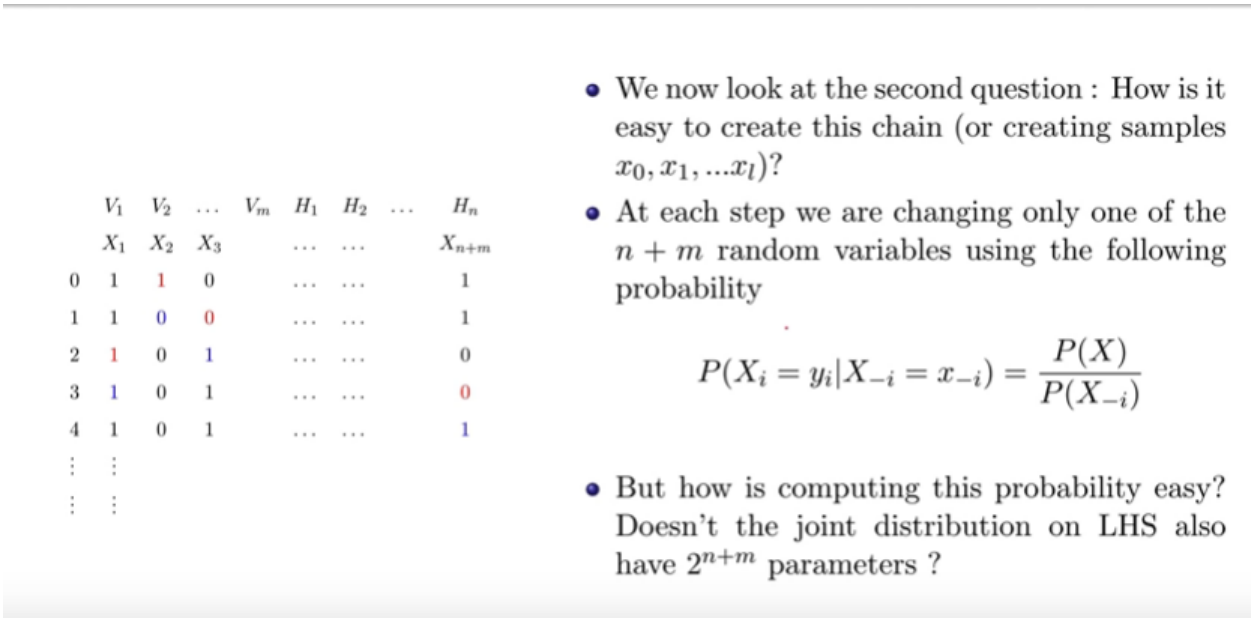
- More formally, we defined T such that

$$p_{\mathbf{xy}} = \begin{cases} q(i)P(y_i|x_{-i}), \text{ if } \exists i \in \mathbf{X} \text{ so that } \forall v \in \mathbf{X} \text{ with } v \neq i, x_v = y_v \\ 0, \text{ otherwise} \end{cases}$$

- where $q(i)$ is the probability that X_i is the random variable whose value transitions while the value of X_{-i} remains the same
- The second term $P(X_i = y_i|X_{-i})$ essentially tells us that given the value of the remaining random variable what is the probability of X_i taking on a certain value
- With that we have answered the first question "What is the transition matrix T?" (It is a very sparse matrix allowing only certain transitions)

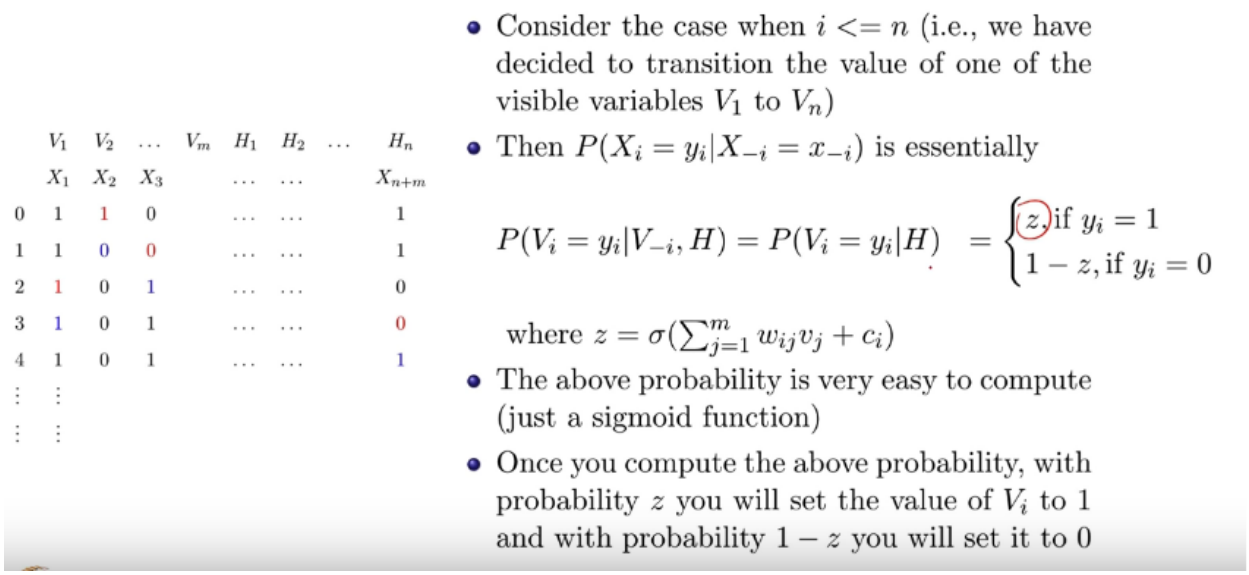
I've answered the first question, what is the transition, is there some confusion here? Sure. Okay? Have answered the first question, what is the transition matrix and the answer is it's a very, very sparse matrix. So, I don't need to define to raise to n plus m, cost to raise to n plus m values, I need to define only a few of those and I have given you the form for that is that clear. Okay?

Refer Slide Time :(21: 48)



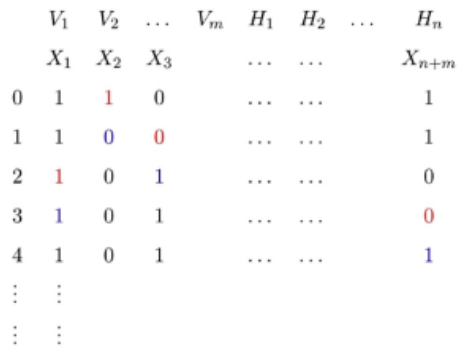
Now, we look at the second question: which was how is it easy to create this chain? That means at every point, what's the computation that I need to do? I need to fix one variable that's easy, once I fix this variable I need to compute that probability value. Okay? So, what's the probability value that, I need to compute this one. That the ayath variable takes on the value Y I. Right? Given all the other random variables. But, this is how you would compute it. Right? This is how you would compute it, because all you have is a joint distribution. Right? We are always learning the joint distribution. So, you have P of X. So, that's what you have? How many parameters does P of X have? How many parameters does P of X have? How many random variables the stakes have, how many n plus, each of them is binary. So, how many parameters does this have? To raise to, n plus M that means you have completely forgotten the lecture on RBMS. But, it's ok. So, 2 raise to n plus M Okay? So, this is not easy. Right? This is again the quantity on LHS, requires you to give me2 raise to n plus M values. So, how is it easy? What do we do to make a Joint Distribution easy, factorize it did we factorize it, did we factorize it for RBMS, yes, ok. If you have forgotten it, let's see. Okay? This is not actually hard and I will tell you, why this is not hard.

Refer Slide Time :(23: 21)



So, consider the case when I is less than equal to n that means, the random variable that you have picked, to change is one of the visible random variables. Right? You have n plus M values, the first n are the visible variables, again goofed up. Okay? N and M we have confused throat. Right? So, it's one of the visible variables. Okay? Let's just focus on that. Okay? So, if I is less than equal to M that means I have decided to change one of the visible random variables. So, if I have done that, then this is the prodded at I'm interested in, which is the same as saying, I'm Interested in the probability distribution. Right? That what is the probability that the I ate visible variable takes on a certain value, given the remaining visible variables and the remaining hidden variables. Okay? But, this actually simplifies because, where are the independence assumption and I just need to compute this probability of the ayath, visible variable taking on a certain valuable value, given all the hidden random variables and that is equal to some Z, again is it easy to compute the Z, did you derive this in your assignment, did we derive this in the last class, what was this? Why are RBMS, neural network, what is this quantity? Sigmoid of, gradient, who said gradient? random, sigmoid of summation W IJ, VJ plus CI. I mean even if you don't remember, you remember that there was a neat closed form solution to this, is it hard to compute this, no it's just a simple vector multiplication. Right? Is just a sum of W IJ, V IJ plus. So, that means I can compute the probability of, the visible variable taking on the value 0 or the value 1. Okay? Because that's those are the only two values possible and I can compute this very, easily? So, this was the term that, I was worried about and I was worried about whether I can compute this efficiently, on the previous slide, Is cared you by saying oh this has to raise to n plus m parameters. But, on this slide I'm, making a case that we have already simplified this in the previous lecture, where we had said that, this just boils, down to computing a sigmoid function. So, I have this sigmoid function and I will send the value to 1, with the probability Z and the value to 0, with the probability 1 minus say. How do I do this in practice? How do I do this in practice? So, I computed the sigmoid, suppose that gave me the value point 2.Okay? That means that was point 2 that means I need to set this, to point 2 with a probability 1 and 0 with a probability0.8. How will I do this everyone? You will you will pick up a random variable, a random value between, zero to one sample from a, uniform distribution. If the value is less than 0.2, you will set it to one, if the value is greater than0.2 you'll set it to zero. Right? So, this is straightforward, there is no problem in doing this. Okay?

Refer Slide Time :(26: 41)



	V_1	V_2	...	V_m	H_1	H_2	...	H_n
	X_1	X_2	X_3		...	...		X_{n+m}
0	1	1	0		...	...		1
1	1	0	0		...	...		1
2	1	0	1		...	...		0
3	1	0	1		...	...		0
4	1	0	1		...	...		1
⋮	⋮							
⋮	⋮							

- So essentially at every time step you sample a i from a uniform distribution (q_i)
- And then sample a value of $V_i \in \{0, 1\}$ using the distribution $Bernoulli(z)$
- Both these computations are easy
- Hence it is easy to create this chain starting from any x_0

So, essentially at every time step is the same as saying that, we sample a I, from a uniform distribution Q I which is very easy to do, once we have done that, we sample from the Bernoulli distribution, with mean Z and that's exactly, what we are doing with, it's a Bernoulli distribution, because it have binary variables which can take on only values 0 to 1, the mean of the distribution is Z that means the probability of heads or the probability of taking on the value 1 is Z. So, you just need to draw from this distribution and I told you the exact procedure of how you will draw from this distribution, you'll actually sample a number from a uniform distribution between 0 to 1, how the value is less than Z, you will set it to 1, if the value is greater than say - you'll set it to 0. Everyone is clear; everyone sees that this is actually easy to do, easy to do. So, what is the question that I have answered now? It is easy to draw samples from this chain, starting from any value X 0. I have X 0; I know how to transfer to X 1, X 2 and so on, efficiently. Right? The computations involved are very simple, is that clear.

Refer Slide Time :(27: 47)

- Okay, finally let's look at the third question: How do we show that the stationary distribution is $P(X)$ (where $X = V, H$)
- To prove this we will refer to the following Theorem:

Detailed Balance Condition

To show that a distribution π is a stationary distribution for a Markov Chain described by the transition probabilities p_{xy}, $x, y \in \Omega$, it is sufficient to show that $\forall x, y \in \Omega$, the following condition holds:

$$\pi(x)p_{xy} = \pi(x)p_{yx}$$

- Let us revisit what p_{xy} is and what π is

So, now we have answered the first two questions and now, what does the third question: that we need to answer that the stationary distribution, is P of X. Right? We need to show that the stationary distribution of this particular Markov chain, with the transition probabilities defined by that, T which, I just very clearly explained to you, is going to be, P of X that's what I need to show. Okay? How many of you are completely lost at this point? How many of a following everything at this point? Okay? The safe question is 80%, how many of you are falling eighty percent at this point? Please, raise your hands high up. Okay? that's good, 20% of course you will go back and read the slide that's why, the slides are there. So, to prove this, you're, going to rely on the following theorem: which is something known as a detailed balanced condition, what this says is that? If I want to show that a certain distribution pi, is a stationary distribution for a Markov chain, with transition probabilities P of X Y. Right? So, just be aware that I'm calling, 'T' of 'X Y' as 'P' of 'X Y'. Right? With these transition probabilities and we actually had defined P of X Y on that, slide which shall not be named, it's sufficient to show that, for all X Y belonging to my sample space the following condition holds. Okay? I know it's very hard to understand, what this condition is? Why we need to prove it and so on. So, I am NOT going to prove this theorem: we are going to take this theorem for granted that the detailed balance condition is sufficient, to show that, the distribution pi is a strange stationary distribution for the given Markov chain. Right? We are going to have faith in this theorem and we'll show that, for our particular case this condition holds, hence the distribution that we care about is the stationary distribution. So, you see where we are headed now, they are going to rely on this theorem and prove what we need to prove. Okay? So, now what are P of X comma Y and what is pi, these are the two things that will first define clearly. Right? Because these are the two things that we need here.

Refer Slide Time :(30: 03)

- Recall that $p_{\boldsymbol{xy}}$ is given by

$$p_{\mathbf{xy}} = \begin{cases} q(i)\pi(y_i|\boldsymbol{x}_{-i}), \text{ if } \exists i \in \mathbf{V} \text{ so that } \forall v \in \mathbf{V} \text{ with } v \neq i, x_v = y_v \\ 0, \text{ otherwise} \end{cases}$$

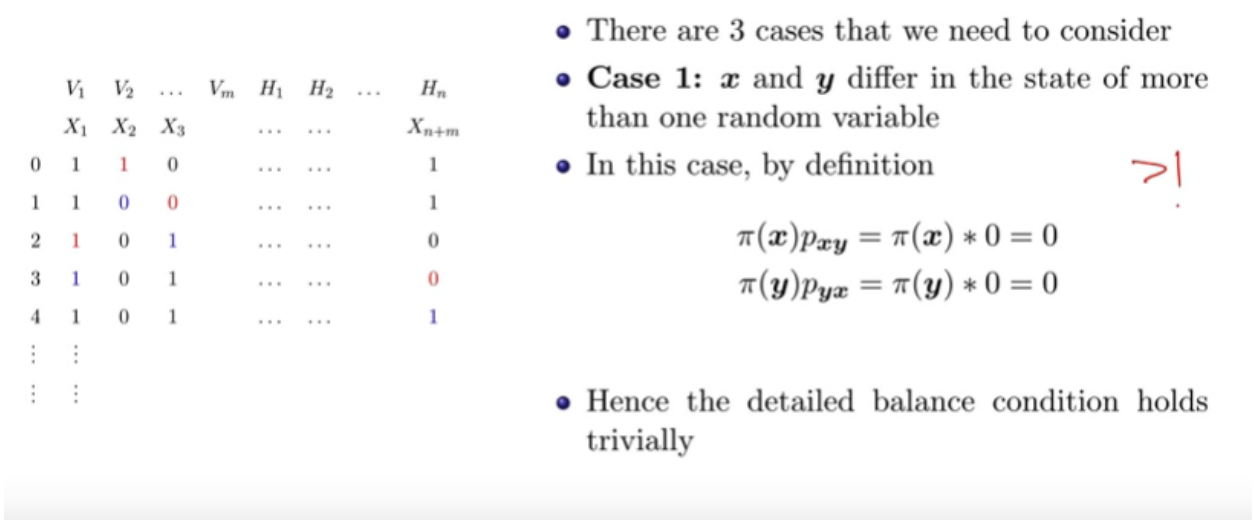
- For consistency of notation we will denote $P(X)$ i.e., $P(V, H)$ as $\pi(X)$
- Further, as shorthand we will refer to $\pi(X = \boldsymbol{x})$ as $\pi(\boldsymbol{x})$
- Thus, to prove that $P(X)$, i.e., $\pi(X)$ is the stationary distribution for our Markov Chain we need to prove that

$$\pi(\boldsymbol{x})p_{\boldsymbol{xy}} = \pi(\boldsymbol{y})p_{\boldsymbol{yx}} \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \{0,1\}^{m+n}$$

$P(X_i = x \mid X_{i-1} = y)$

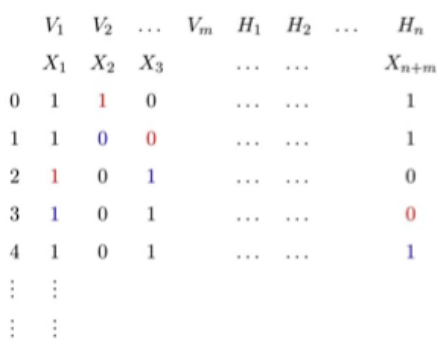
So, first P of X Y. Okay? That's what this is? I allow only transitions of one particular random variable and for that random variable I define it, I was calling it, 'P' instead of pi, but now, on I'm going to change this. So, we are going to refer to P of X, which is actually P of V comma H, as pi of X. Right? Just to be consistent with the notation. Okay? Avian fine with that just P has become pi. Okay? And for shorthand, instead of saying property of capital X, taking on the value small X, I'll just call it as, 'PI' of 'X'. So, when I say PI of a small letter, it means that the capital letter, taking on the small letter well whatever that means. Okay? Okay? Capital X taking on the value small X is that fine I'm just going to refer to that as shorthand and similarly when I say, X Y, it actually means transitioning from. Right? That's what it means, oh is that Okay? Right? So, you know what PI is and you know what P is and now, we need to prove that this detailed balanced condition holds. Okay?

Refer Slide Time :(31: 25)



Now, there are three possible cases, first let's agree that these are the three cases that cover all possible cases. The first case is, when x and y actually differ in more than, two values. Okay? That's one case. The second case is, when x and y do not, differ in any value. Okay? That means they're exactly equal. At the third cases, when x and y differ only in one value, are there any more cases possible, no it could be visible or hidden. So, the same argument which I gave for visible that if it was, if it was less than n, then it's easy to compute, the same argument holds for if it was between n plus M. So, I'm saying there are three possible cases, one is when x and y differ and zero values that means they're, exactly the same, the other cases when they differ in exactly one value and the third cases when they differ in sorry, more than one values, my discovers everything, there is no other case possible here, is that clear. Okay? So, if I prove that for all these three cases, the detail balance condition holds, then I'm done well. Okay? So, let's start with the easy case one: which is x and y differ in more than one state that clear, more than one random variable, sorry. So, in this case, by definition you want to prove that PI X, PX Y, is equal to PI X into zero, which is zero and the other way around also it's zero. So, trivially the detailed balance condition holds, the case one was very easy. Right? The greater than one case, was very easy, when the state's x and y differ in more than one random variable, it's trivial.

Refer Slide Time :(33: 15)



	V_1	V_2	…	V_m	H_1	H_2	…	H_n
	X_1	X_2	X_3		…	…		X_{n+m}
0	1	1	0		…	…		1
1	1	0	0		…	…		1
2	1	0	1		…	…		0
3	1	0	1		…	…		0
4	1	0	1		…	…		1
⋮	⋮							
⋮	⋮							

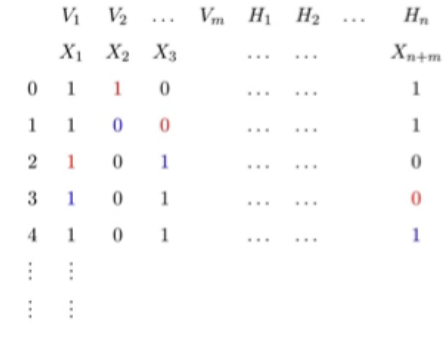
- There are 3 cases that we need to consider
- **Case 2:** $\boldsymbol{x}$ and $\boldsymbol{y}$ are equal (i.e., they do not differ in the state of any random variable)
- In this case, by definition

$$\pi(\boldsymbol{x})p_{\boldsymbol{xy}} = \pi(\boldsymbol{x})p_{\boldsymbol{xx}}$$
$$\pi(\boldsymbol{y})p_{\boldsymbol{yx}} = \pi(\boldsymbol{x})p_{\boldsymbol{xx}}$$

- Hence the detailed balance condition holds trivially

Let's look at the other case, when X is equal to Y that means the case when they differ in 0 random variables. In this case, again it holds trivially, PI X into PX Y is the same as PI X into P xx and PI Y into p IX is again the same as PI X into PX X. Right? So, again trivially the detailed balance condition holds. So, for clear. Okay?

Refer Slide Time :(33: 38)



	V_1	V_2	…	V_m	H_1	H_2	…	H_n
	X_1	X_2	X_3		…	…		X_{n+m}
0	1	1	0		…	…		1
1	1	0	0		…	…		1
2	1	0	1		…	…		0
3	1	0	1		…	…		0
4	1	0	1		…	…		1
⋮	⋮							
⋮	⋮							

- There are 3 cases that we need to consider
- **Case 3:** $\boldsymbol{x}$ and $\boldsymbol{y}$ differ in the state of only one random variable
- In this case, by definition

$$\begin{aligned}\pi(\boldsymbol{x})p_{\boldsymbol{xy}} &= \pi(\boldsymbol{x})q(i)\pi(y_i|\boldsymbol{x}_{-i}) \\ &= q(i)\pi(x_i,\boldsymbol{x}_{-i})\frac{\pi(y_i,\boldsymbol{x}_{-i})}{\pi(\boldsymbol{x}_{-i})} \\ &= \pi(y_i,\boldsymbol{x}_{-i})q(i)\frac{\pi(x_i,\boldsymbol{x}_{-i})}{\pi(\boldsymbol{x}_{-i})} \\ &= \pi(\boldsymbol{y})q(i)\pi(x_i|x_{-i}) \\ &= \pi(\boldsymbol{y})p_{\boldsymbol{yx}}\end{aligned}$$

- Hence the detailed balance condition holds

Now, we come to the case, where x and y differ in exactly one random variable. Okay? So, in that case, by definition we this is our LHS. Right? So, we want to show that this LHS is actually equal to the following RHS; this is what we need to show. Okay? So, now by definition the second quantity, P of X Y is actually equal to this, avian fine with this, do you guys want to break? So, anytime I get an answer immediately, it's already 6:30, you have a long night ahead of us, 39 out of 61. Right? So, Okay? Let me just finish this part so, it's logically done and then we'll take a break. Okay? So, I have just substitute the value of P XY, which was defined on the previous slides, is that Okay? Fine lunch is going to do some very simple trickery. So, let's again be sure that, this is the value that, one of the random variables takes on, that is the ayath random variable takes all and this, is a vector because this is the value that the remaining random variables take on. Okay? Is that fine. Now, this thing, I can write it as, this thing, a conditional distribution is joint over marginal, how many for get this? Please raise your hands up in height. Okay? And this guy, I'm just going to split it into X I and X minus 1, is that Okay? So, this is the same as saying that X I is equal to small X I and the remaining random variables equal to X minus I is that Okay? And I've just split it into two parts; I'm just rearranging the terms. So, I have taken this and put it here and I've taken this and put it here, however attach this Q I here, just are arrangement of the terms. Okay? Now, what is this quantity? This is X I given, X minus I and what is this actually? PI by PI X, PI Z by what is that this quantity everyone? Anyone who has a doubt about it? Please raise your hands and I do not understand anything the rest of the course. Okay? So, this is PI y and this is X I given X I minus one. Now, what is this combined? Actually there is one more step here, this X minus I, I can just call it even why minus I, write because for the - I variables x and y are same. Okay? So, now what's this circled quantity? P of Y X. Okay? So, starting from the LHS we have come to the RHS. So, we have shown that, the detailed balance condition holds for, all the three cases, which were possible, hence PI is the stationary distribution of Earth Jupiter, what? The stationary distribution of, of this Markov chain. Right?

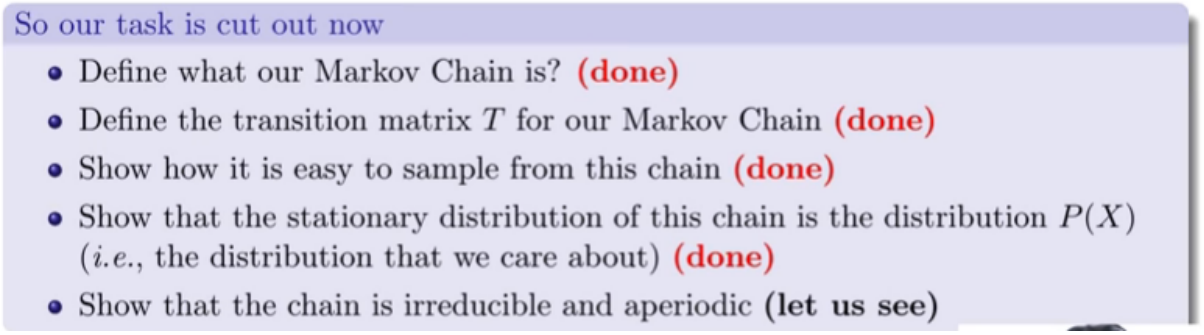
Refer Slide Time :(37: 18)

	V_1	V_2	...	V_m	H_1	H_2	...	H_n
	X_1	X_2	X_3		...	...		X_{n+m}
0	1	1	0		...	...		1
1	1	0	0		...	...		1
2	1	0	1		...	...		0
3	1	0	1		...	...		0
4	1	0	1		...	...		1
⋮	⋮							
⋮	⋮							

- Thus we have proved that the detailed balance condition holds in all the 3 cases
- **Case 1:** $\boldsymbol{x}$ and $\boldsymbol{y}$ differ in the state of more than one random variable
- **Case 2:** $\boldsymbol{x}$ and $\boldsymbol{y}$ are equal (i.e., they do not differ in the state of any random variable)
- **Case 3:** $\boldsymbol{x}$ and $\boldsymbol{y}$ differ in the state of only one random variable

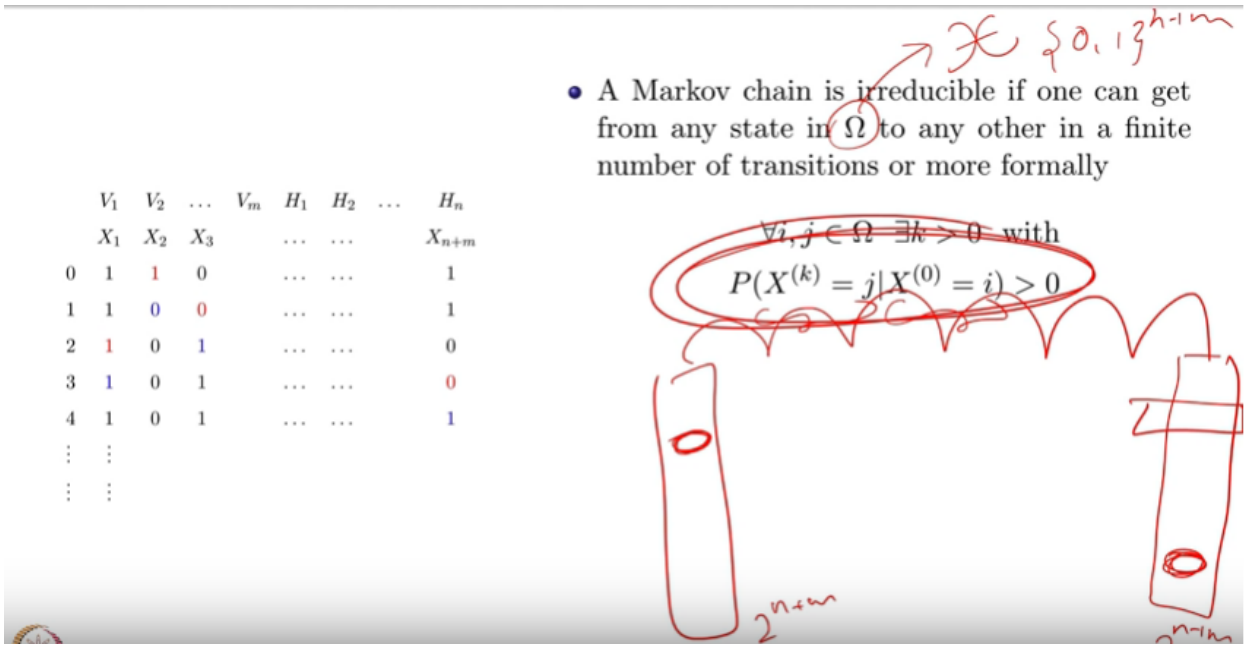
So, we have setup the Markov chain, in such a way that we have been able to prove that the detailed balance condition holds for all these three cases. This is the case 1, this is the case 0 and this is the case greater than1. Right? For all these cases it holds. So, that means if we run this Markov chain, for enough number of time steps, we are going to reach the stationary distribution that we care about if you keep sampling, values from this Markov chain, eventually we will get values, which are same as if they had come from the distribution that we care about. Right? And even after we reach the stationary distribution, the sampling process remains the same. Right? You're going to use the same sampling process, which was set, one value, keep everything as the same and just change this value. But, now with this simple procedure actually, have started getting, actually of starting getting samples, which come from your Joint Distribution that you care about. So, you see the overall trick. Right? You could set up a chain, which is.

Refer Slide Time :(38: 18)



So, these were the things that we cared about. Right? Define what a Markov chain is? Have you done that? Right? How we define the transition matrix done. Okay? Is it easy to sample from this chain, done, have you shown that the stationary distribution of this chain is the distribution that we care about done, have you shown that the chain is a reducible and a periodic done, you kind of showed everything that we had to prove, to what effect we are proving all this is still not clear. But, we'll get there but, this is one thing which we have not shown, remember the statement of the theorem had that this, chain has to be a periodic and irreducible, I have not proved that, I am not even defined what a periodic is and what irreducible is. So, we'll quickly, take a look at it it's, it's the easy part of this lecture.

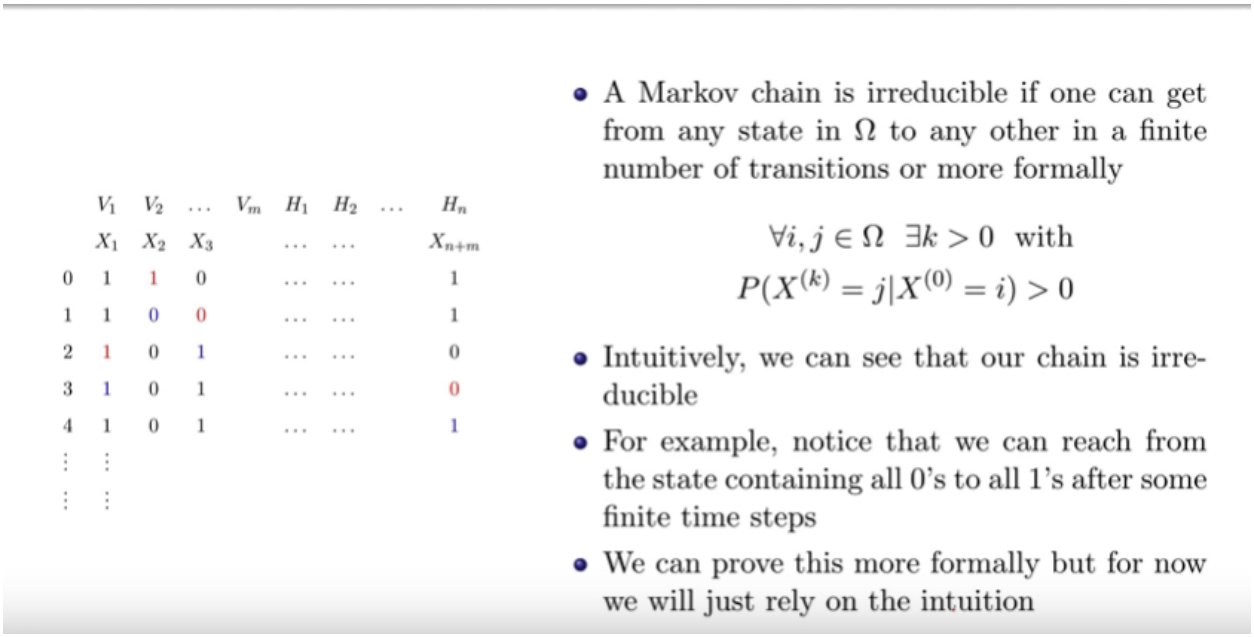
Refer Slide Time :(39: 07)



So, we'll just do that quickly. So, first I look at the definition of irreducible. So, a Markov chain, is irreducible, if when one can get from any state, in Omega Rights, Omega is your entire state space, this was earlier defined using that stylish X or its 0 to1 raise to n plus m in our case. So, in your state, in your state space, if it is possible to reach any state, starting at any initial state. Right? So, you have, to raise to n plus M possibilities for this and you have to raise to n plus M possibilities for this, what I'm saying is that? If I want to reach from any state to any other state and there exists a value K, such that after that many steps, there will be some probability that I am going to be able to reach this other state. Right? So, in other words the, other way of looking at it is that, it is not the case that, if you have one of these to raise to n plus M values that you start from, the remaining 2 raise to n plus n minus 1values are reachable, from here it's not that they're a particular state, which is not reachable from there. So, it's not the case that, the probability of reaching a particular state, starting from any other state is 0 that's, what it means irreducible because the chain will continue, is that fine, is the definition clear, what irreducible means? Please raise your hands if it's clear. So, it basically just means that, I know that, after some K steps at least, there is a finite probability of, reaching any state starting from any other state. So, this holds for all I comma J belonging to Omega. So, what about our Markov chain is it irreducible? So, a Markov chain is irreducible. So, you have these two raise to n plus M values that your States can take. Right? So, let's see I started with one of these values. Okay? And now, the question that I'm interested in is that I'm going to run this Markov chain for many time steps. Okay? Now, the question that I'm asking is that is it possible and if I take any other state, from this to raise to n plus M States, can I reach that state, even if I run it and I'm not putting any restriction on Carol, I'll keep running it. Okay? Is it possible that at some time step K, arbitrary time steps K, all these to raise to n plus M values, are reachable, irrespective of where I start from. Right? So, if there is a break in between, if there is a case that, I cannot reach one of these values, then it would mean that this probability, is equal to 0 for all case, there does not exist a single K, for which this probability is greater than 0that means no matter how, long you run the chain, starting from a particular state, you cannot reach this other state. If that happens, the chain if that does not happen, then

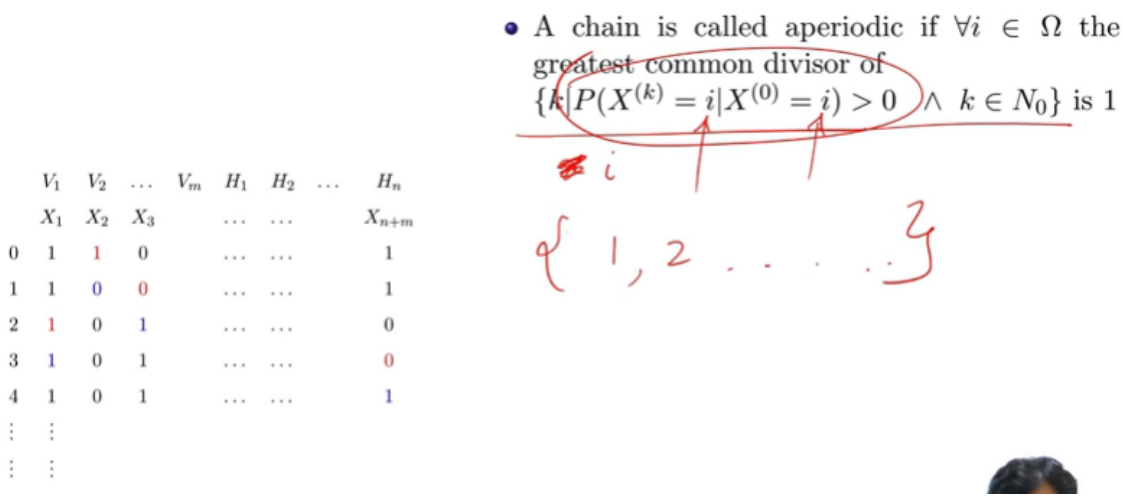
the chain is called, ‘Irreducible’ that means given any state, starting state, you can reach all the other states after some number of time steps and that's what a reducible means.

Refer Slide Time :(42: 06)



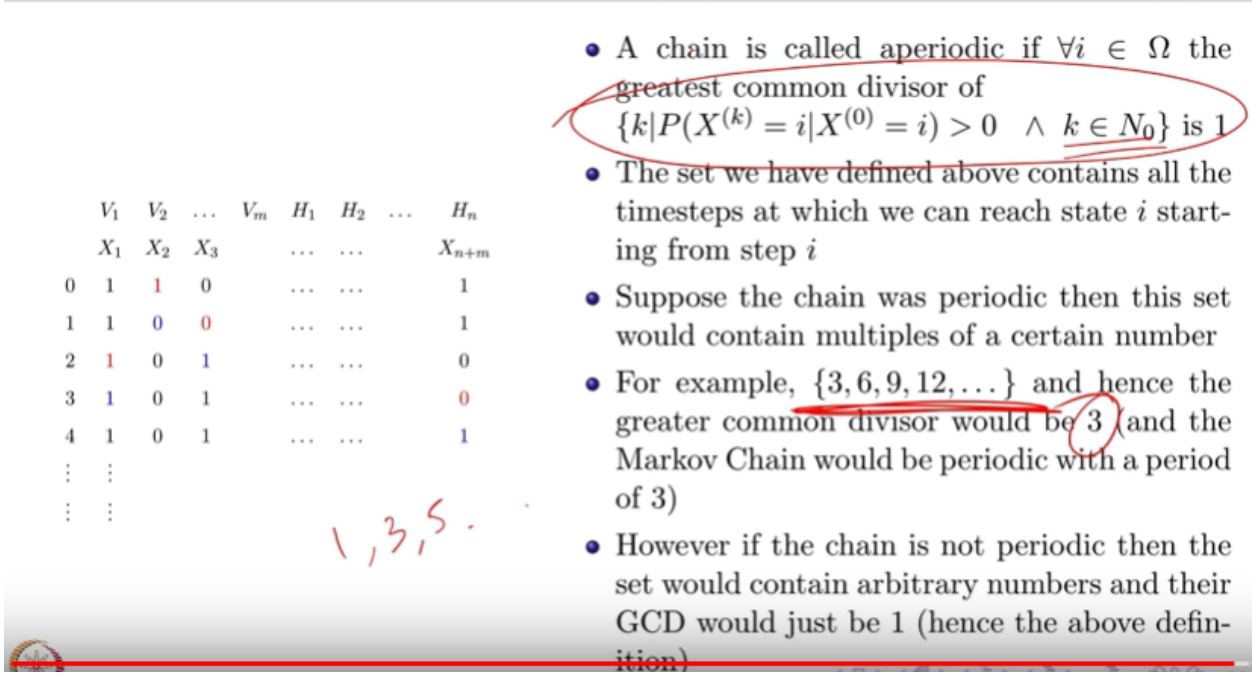
So, the Markov chain that we had is it irreducible, yes. Right? And you can take the very simple case; the most difficult transition would be starting from all zeros and going to all ones. Right? Because you are allowed to only change one value at a time. Right? Even in this case, it's possible to reach from all zeros, to all ones and if you could do that, then it's possible to reach any state from anywhere and this is a very intuitive explanation, is that clear, everyone gets this, should be straightforward to see that the chain is, irreducible. Okay? So, we can prove this more formally, but we are not going to do this, we will just live with the intuition that, this chain is irreducible hence 1 red mark which we had in a theorem: that it was true only for irreducible Markov chain, we don't need to worry about this.

Refer Slide Time :(42: 53)



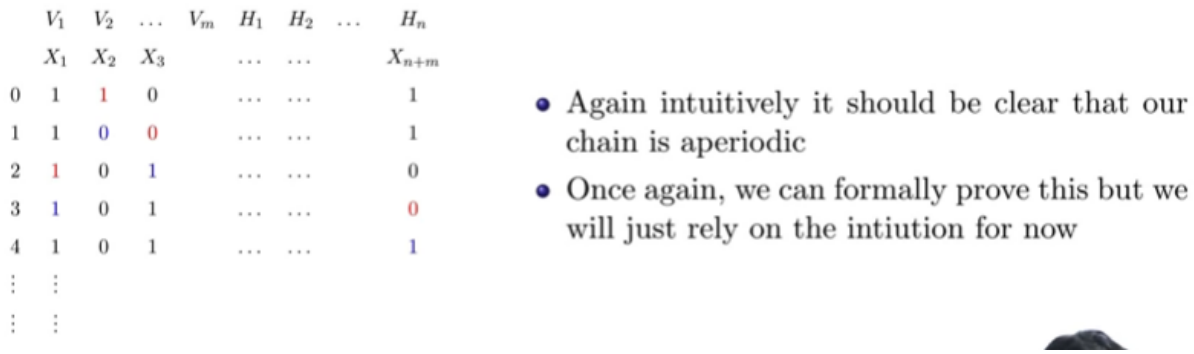
Now, the other thing is even more cryptically defined. So, a chain is called, 'A periodic'. If the greatest common divisor of this set is 1, what is the set? What does this set contain? All the time steps at which, starting from the state I at time step 0, I will end up in state I again. Right? Starting from time step state I at time step 0, there is a finite probability, greater than 0 that I'll end up at time step at, at the state I, again at the time step K. So, this is actually a set of the following form that maybe, if I start with some value or some I at time step 0, at time step 1, I can it's possible that I'll stay in the same value. It to it is possible and so on and this is just a collection of those numbers. Okay? So, this set is a collection of all, the time steps, at which starting from one value, there is a finite chance that I can reach that value again. Okay? Now, instead of a periodic, let's talk about periodic, if this was periodic, what would you expect this set to contain? What does periodic mean? Something which happens, periodically, something which happens periodically? Right? So, then what kind of values with this set contain, what kind of values with this set contain? Multiples of some number. Right?

Refer Slide Time :(44: 32)



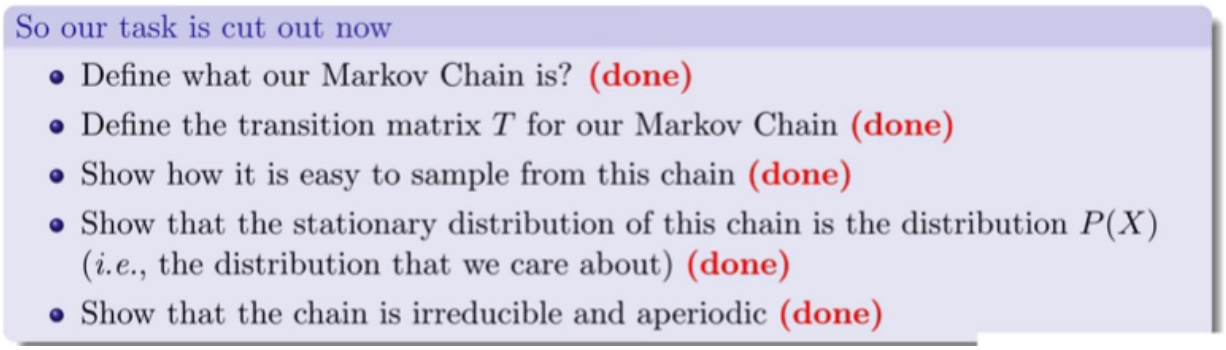
So, it contains so, this is all fine. If it contain these kind of values, a 3, 6, 9, 12that means it's saying that starting from state, from a value, of former state I at time step 0, I can only reach the state I again add these, multiples that means this is periodic. Right? I cannot reach it at anything which is, not a multiple of this period, is that fine. So, that's what periodic means? So, periodic means that this, set would have a greatest common divisor, which is greater than 1. Right? Because it the greatest common divisor would actually be the dash, of the Markov chain, the period of the Markov chain. Right? Now, what a periodic means is that? I will get a set, which does not contain such multiples, it contains some arbitrary values. So, it might contain 1, 3, 5 and so on. So, now if I try to take the greatest common divisor of this set, I will get 1, 1 can be the only greatest common divisor of this set. So, it just means that there is no, there is no pattern to this, there is no period after which this repeats, repeats it can arbitrarily grid, return to the same state, after any number of time steps is that clear. So, it's always helps to understand periodic, then a periodic is just the reverse of that. So, this is what periodic is? So, a period in and if it's periodic, you will have a greatest common divisor, which is not equal to 1, you will have the greatest common divisor as the period of the chain. But, if it's not periodic, then you'll just have the greatest common divisor as 1, because that's, only the device greatest common divisor of all the elements in your set, no strictly speaking no. Right? So, periodic has to be 0, 2, 4, 6, 8 and so on. When you're just changing I so, by shifting then it would be periodic. Right? So, then it goes in that case it would be, K is a natural number, I had asked for a different end, but, ok. K is a natural number. Right? Because it can be time steps, which are greater than 1. Okay? So, I mean just don't, look at this is unfortunately the way of stating, what a periodic is that but don't focus on that, just understand what periodic is and a periodic is just the opposite of that, is that clear. Okay? And that's why that greatest common divisor comes in there.

Refer Slide Time :(46: 42)



Now, intuitively do you think our chain is periodic or a periodic a periodic. Right? So, again we can formally prove this. But, we're just going to rely on the intuition that this is a periodic and more.

Refer Slide Time :(46: 55)



So, now I am done with all the parts of the proof. So, I have set up a Markov chain, which is easy to draw from and I've shown that the stationary distribution of this Markov chain is the distribution that we care about, even when I reach the stationary distribution, I can still, keep following my process of sampling, which is the Markov chain process. But, now after this point, I start getting samples, as if, they were drawn from P of X or PI of X is that clear fine.

Refer Slide Time :(47: 23)

- Okay, so we are now ready to write the full algorithm for training RBMs using Gibbs Sampling

So, I have done what was required.