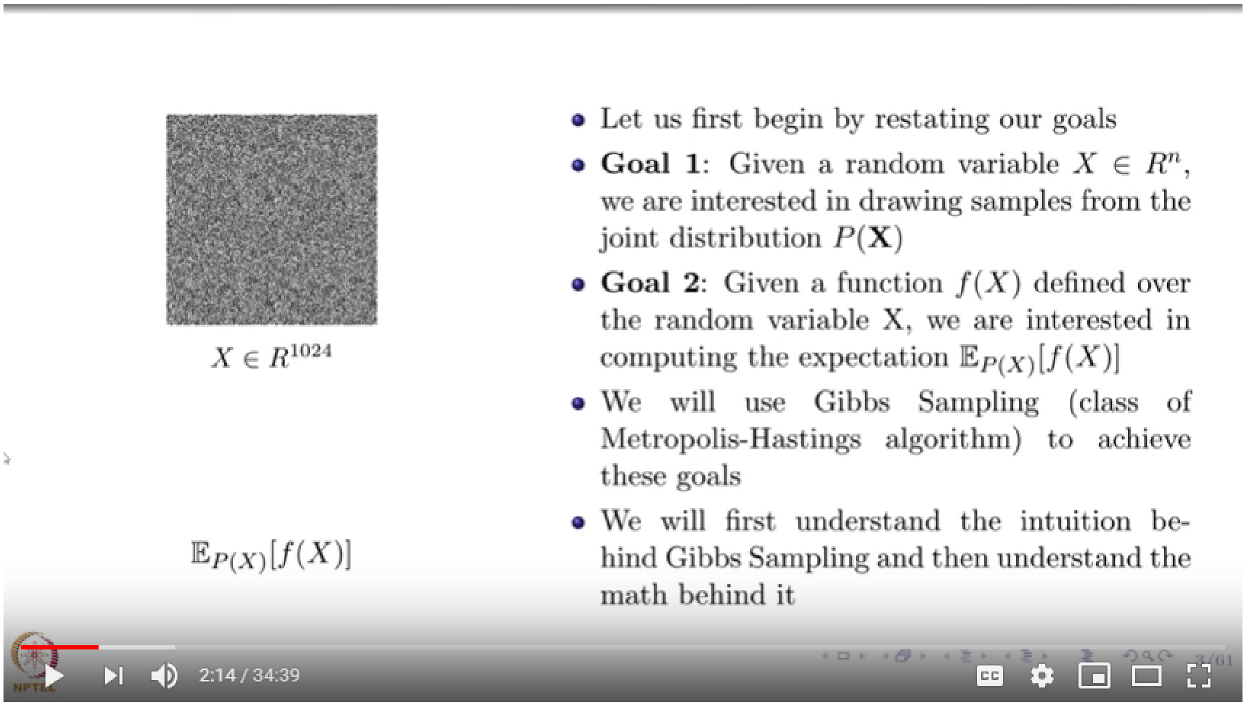
Lecture 19.1
Markov Chains

And today we're going to talk about, Markov chains Gibbs sampling, for training RBMs and then contrastive, divergence for training RBMs. So, it’s a, it's a longest lecture, I think it's also one of the, hardest lectures, just to set the expectations right. And so, that you don’t doze off. There's going to be a lot of material, that we'll cover today and also a lot of math, that will do around, along the way. But I've tried my best to simplify things and I'm pretty sure that if you say stay awake and attentive, you will get most of the stuff. Okay? Okay. So, with that very encouraging note, let’s start the lecture. So, we will start with Markov chains.

Refer slide time :(0:51)



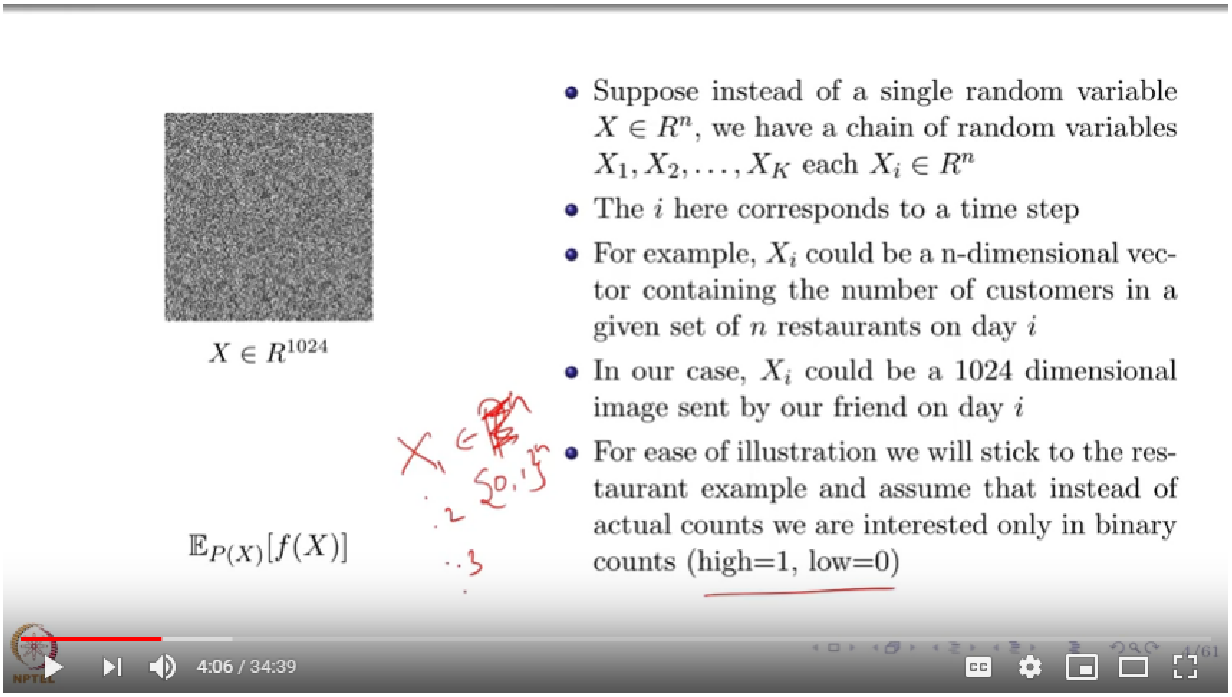
$X \in R^{1024}$

$\mathbb{E}_{P(X)}[f(X)]$

- Let us first begin by restating our goals
- **Goal 1**: Given a random variable $X \in R^n$, we are interested in drawing samples from the joint distribution $P(\mathbf{X})$
- **Goal 2**: Given a function $f(X)$ defined over the random variable X, we are interested in computing the expectation $\mathbb{E}_{P(X)}[f(X)]$
- We will use Gibbs Sampling (class of Metropolis-Hastings algorithm) to achieve these goals
- We will first understand the intuition behind Gibbs Sampling and then understand the math behind it

So, let us first begin by restating, our goals. Right? So, our goal is that we have a random variable, which is a high dimensional random variable. And one thing that we're interested, in is drawing samples, from the joint distribution, from which this random variable has come. Right? So, we call this a joint distribution, because it has, one zero two four random variables, X 1 2 X 10 2 4. I've never seen these guys before. Okay? So, we have this joint distribution, over these N or 1 0 to 4 random variables and we want to draw samples from, this joint distribution and the other thing or a related thing that were interested, in is that given an arbitrary function f of X write a function, of these random variables, we want to compute the expectation of that of this function, under this distribution. Right? And that's exactly the two expectations that we're interested, in those expectations whatever is inside the expectation, you could think of that as a function of this random variable. Right? So, given any arbitrary function, we want to compute this expectation. And of course these goals are related because if you can draw samples, you can empirically compute the expectation and so on, but just to state them clearly, these are the two goals that we are interested in and now, you're going to use as Gibbs sampling, which is a class of metropolises things algorithm, I don’t know why you would care about that. But it’s a class of some famous algorithms and we will use that to achieve these goals. So, first thing that we will do is we'll first understand, the intuition, behind this and then get you to the math behind it. Okay? Okay.

Refer slide time :(02:19)



$X \in R^{1024}$

$\mathbb{E}_{P(X)}[f(X)]$

- Suppose instead of a single random variable $X \in R^n$, we have a chain of random variables $X_1, X_2, \ldots, X_K$ each $X_i \in R^n$
- The i here corresponds to a time step
- For example, X_i could be a n-dimensional vector containing the number of customers in a given set of n restaurants on day i
- In our case, X_i could be a 1024 dimensional image sent by our friend on day i
- For ease of illustration we will stick to the restaurant example and assume that instead of actual counts we are interested only in binary counts (high=1, low=0)

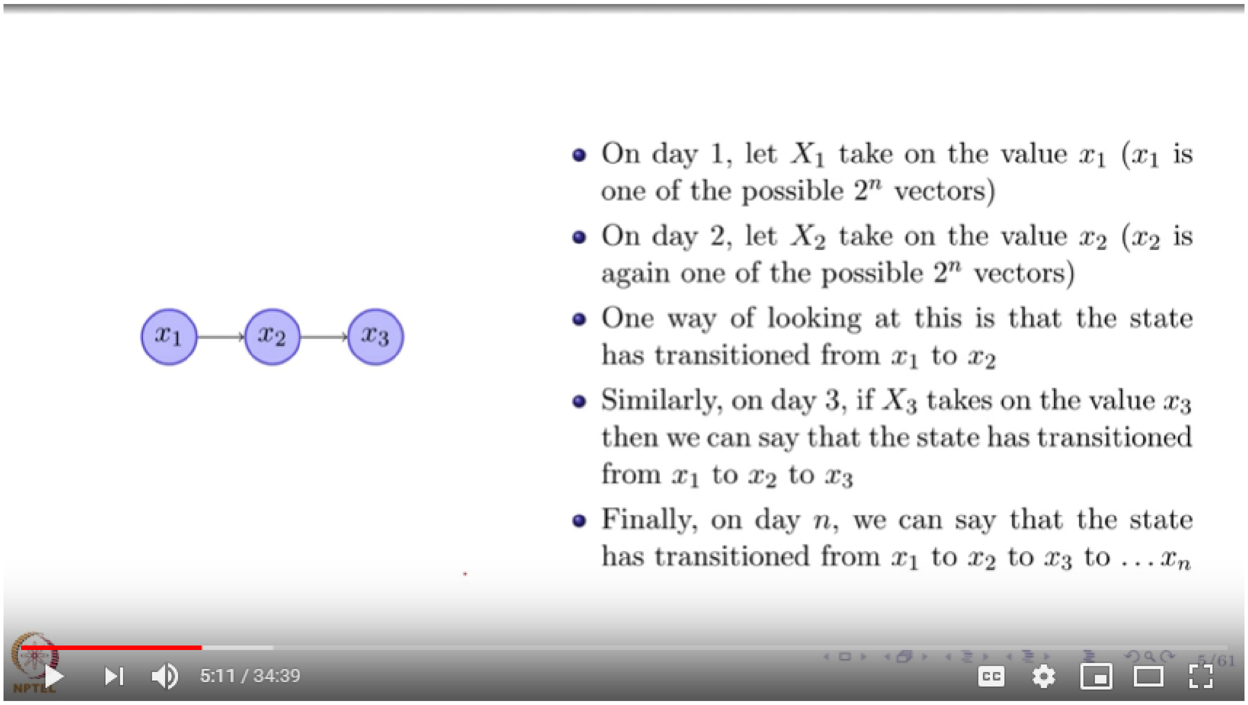
$X_1 \in \cancel{R^n} \{0,1\}^n$

$.2$

$.3$

So, now instead of a single random variable. Right? Which is x belonging to RN, suppose we have a chain of random that means I have X 1, X 2, X K, each of these excise belongs to RN. Right? So, you could think of this as a friend sending us images, on day 1 day, 2 day, 3 day, 4 and so on. So, on day 1 here send us one image, day tow, another image and so on. So, he's seeing this chain of random variables way. And the eye here corresponds to a time step, it's a discrete time step, day 1 day, 2 and so on it's not a continuous time step. And one example which I gave you was images, the other example could be that I have this n dimensional vector, which tells me the number of customers, in a restaurant on day 1, then I have the same vector which tells me the number of customers, on day2 and so, on. Right? So, all of these are X's, but they are also associated with time because it's on, day 1, day 2 and so on. So, I have this chain of random variables. So, it could either be the number of customers in a restaurant or the images or wallpaper sent to us on day 1 day 2and so on. But for this discussion, I'll just stick to the restaurant example that we have this vector, which stores the number of customers, in a restaurant. And we have this for multiple days and just to keep things simple. I will work with discrete variables, what do I mean by that instead of actually having the counts, I'll just have, whether the number of customers was high or low. Okay? Is that fine is a setup clear. So, we have X, which belongs to some space, RN or not R in a sorry, 0 comma 1 raise to N and then we have several of these on day 1 day, 2 day,3 and so on. So, that's the set up that we are working, with so, we have a chain of random variables.

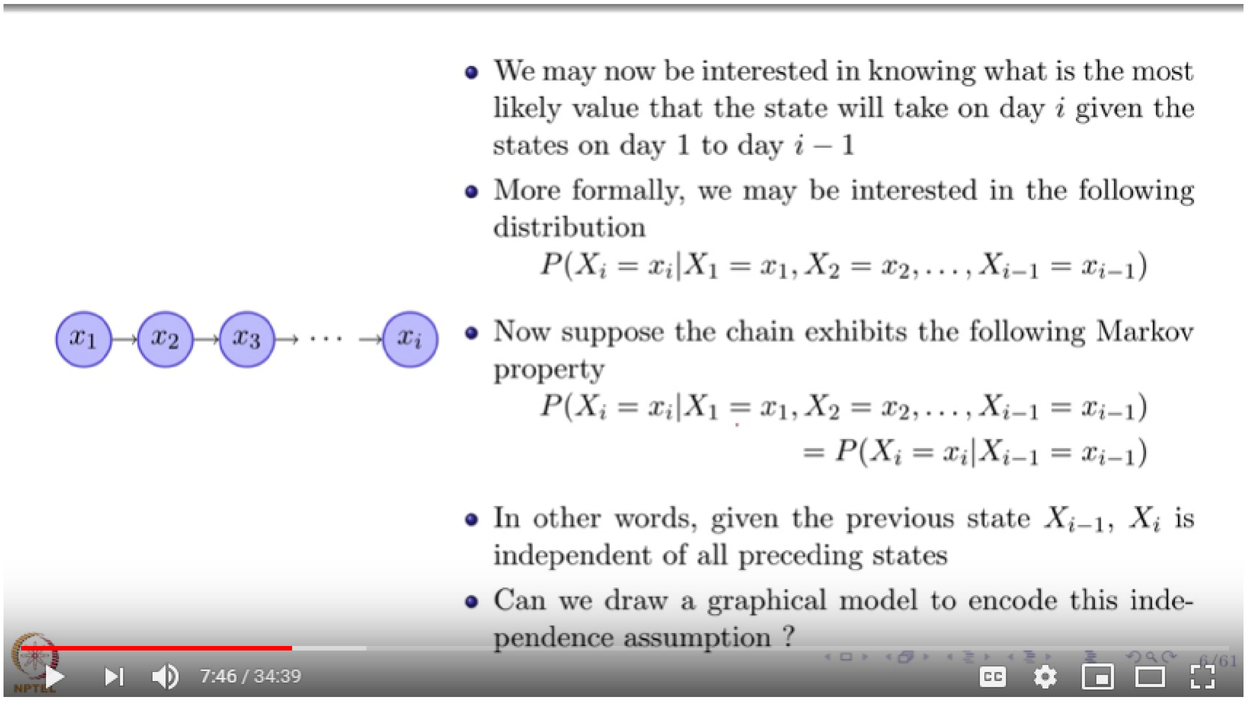
Refer slide time :(04:08)



- On day 1, let X_1 take on the value x_1 (x_1 is one of the possible 2^n vectors)
- On day 2, let X_2 take on the value x_2 (x_2 is again one of the possible 2^n vectors)
- One way of looking at this is that the state has transitioned from x_1 to x_2
- Similarly, on day 3, if X_3 takes on the value x_3 then we can say that the state has transitioned from x_1 to x_2 to x_3
- Finally, on day n, we can say that the state has transitioned from x_1 to x_2 to x_3 to $\ldots x_n$

Now on day 1, let x1take on the value, small x1. Right? And this small x1 is one of the 2 raise to n possible values. Right? So, remember that X can take 2 raise to n possible values, with high low for each of the N restaurants that we have. Right? On day 2let x2 take on the value X small X 2, which is again one of the 2 raise to n possible values that it can take on. Now one way of looking at it is that, the state of the random variable has transitioned fromx1 to x2 from day 1 to day 2 that's a fair way of saying this and you are looking at this random variable the semantics of the random variable, remains the same across time steps, it's the number of customers. But just the state has changed from x1, small X 1 to small X2. And now on day 3, if I assume thatthere are X 3 customers, then I can say that the status transition from X 1 to X2 to X 3 and in general on day n, I can think of it that starting from some small X 1, on day 1 the state has transitioned through x2 x3 up to XN. Right? Okay?

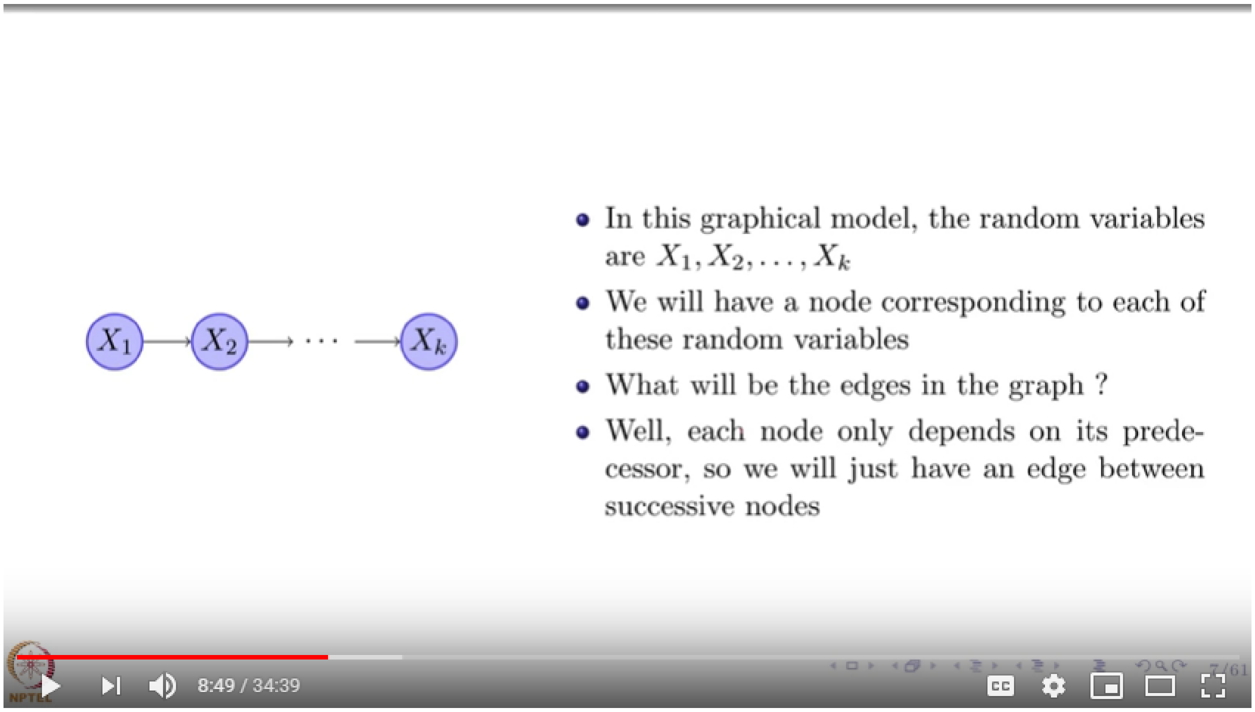
Refer slide time :(05:11)



- We may now be interested in knowing what is the most likely value that the state will take on day i given the states on day 1 to day $i-1$
- More formally, we may be interested in the following distribution
  $$P(X_i = x_i | X_1 = x_1, X_2 = x_2, \ldots, X_{i-1} = x_{i-1})$$
- Now suppose the chain exhibits the following Markov property
  $$P(X_i = x_i | X_1 = x_1, X_2 = x_2, \ldots, X_{i-1} = x_{i-1}) = P(X_i = x_i | X_{i-1} = x_{i-1})$$
- In other words, given the previous state X_{i-1}, X_i is independent of all preceding states
- Can we draw a graphical model to encode this independence assumption ?

So, we are now started talking in terms of states that the random variable, can take and we are transitioning between these states. Now an interesting question, to ask would be what is the most likely value that the state, will take, on day I, given the states, from day 1 today I minus 1. Right? Can you think of an interesting application of this, a financially lucrative application of this stock market. Right? So, you know what the stock price has been over the past, hundred days and you would like to know what value it could take on the next, day I know it's assumed discrete again high or low is what you are interested and not the exact value of the stock. Right? Now more formally what we are interested, in is that we are interested in the random variable X I, each variable in the chain is a random variable, taking on the value small X I, given the states of all the, other random variables that you have seen previously it. So, x1 equal to small X 1 X 2 equal to small X 2 and so on. Right? That's the question that we are interested in, now suppose, the chain exhibits the following Markov property. Okay? So, I'm calling this a,' Markov Property', because I'm assuming that given the previous state, the current state is independent, of all the other states. Right? So, I only care about, what the situation was on day I minus 1, once I know that I can determine what the situation is going to be on day I and I don't care about, what happened from day1 to day I minus 2, this is an assumption again, this is something that we are assuming that the chain exhibits this property. And this ties to all the model assumptions that we have been making so, far in the course so, we're just making an assumption that this is how our chain behaves. Right? So, this is exactly what I said, can you draw a graphical model, to represent this situation. So, the moment I ask you to draw a graphical model, this is the first question that she should think about, for the first, I mean such a trivial question that you'll not even think about, what are the? What other? Random variables in your distribution. Right? So, what are the random variables in your distribution, x1, x2, x3 and so, on the capital x I are your random variables. Okay? So, that's out of the way. So, now let's come to the graph what is going to be, what are going to be the nodes in the graph?

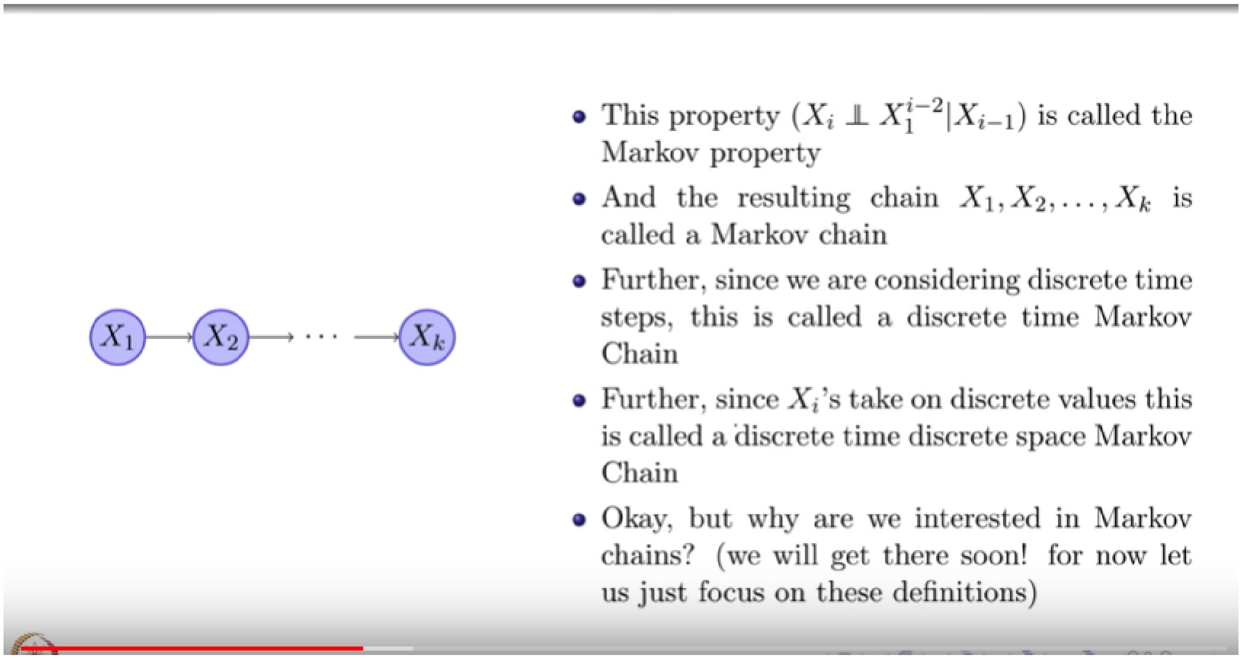
Refer slide time :(07:51)



- In this graphical model, the random variables are $X_1, X_2, \ldots, X_k$
- We will have a node corresponding to each of these random variables
- What will be the edges in the graph ?
- Well, each node only depends on its predecessor, so we will just have an edge between successive nodes

So, you'll have one node, for every random variable that you have so, these are the nodes in the graph. Now what are going to be the edges in the graph, I think everyone should be able to answer that and of course assume the Markov property, what are going to be the edges? The edges indicate what? Dependencies and what do you know about Bayesian networks, given the dash you are independent of the dashes. We are the parents, you are independent of the non-citizens. So, you know what you are independent of given what? So, now what should be, the parent and I'm asking it very cryptically, but I'm assuming we have done enough of this to be able to answer it. So, now what would be the parent of each node, the previous node, given the previous node, it is independent of all the other nodes. So, you have a very simple, Markov Bayesian network, where you just have these dependencies, between the previous guy, to the current guy. Okay? That's very straightforward. Okay?

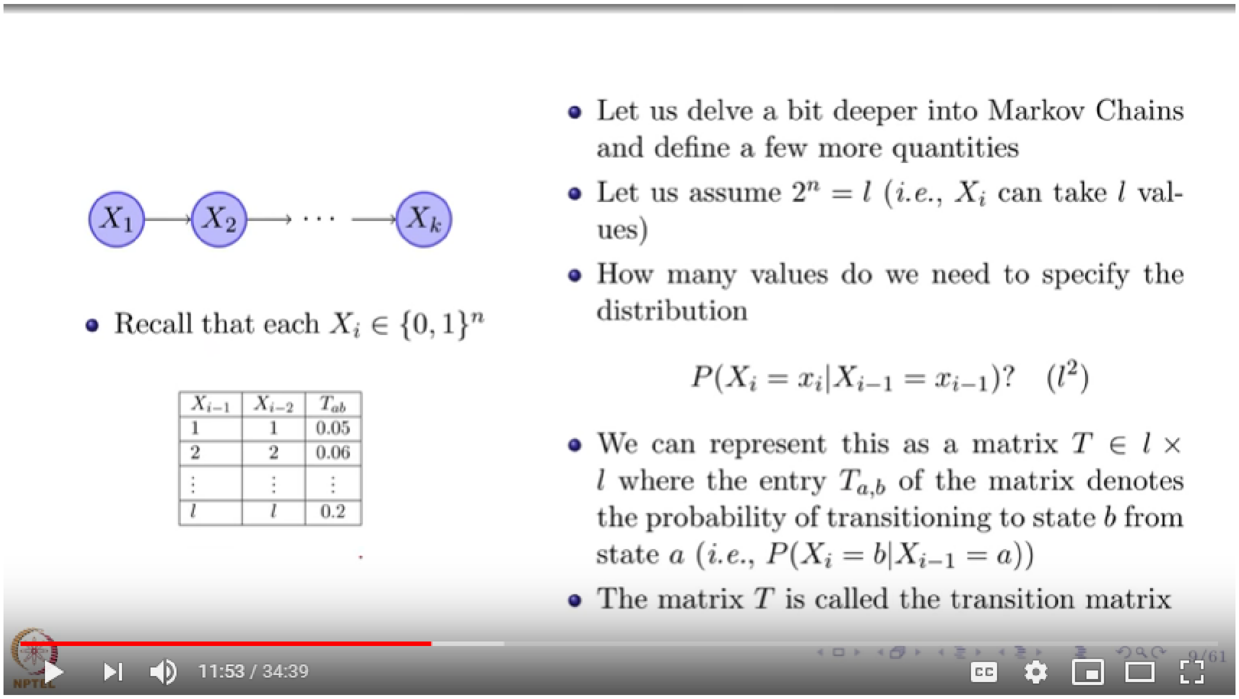
Refer slide time :(08:53)



- This property ($X_i \perp\!\!\!\perp X_1^{i-2} | X_{i-1}$) is called the Markov property
- And the resulting chain $X_1, X_2, \ldots, X_k$ is called a Markov chain
- Further, since we are considering discrete time steps, this is called a discrete time Markov Chain
- Further, since X_i's take on discrete values this is called a discrete time discrete space Markov Chain
- Okay, but why are we interested in Markov chains? (we will get there soon! for now let us just focus on these definitions)

So, this property, where X I, is independent of X 1 to I minus 2 given X I minus 1 is called the,' Markov Property'. And the resulting chain, X 1, X 2 up to XK, is called the,' Marko Chain'. Okay? Just definitions, further since we are considering this cream, discrete time step, so, it our time steps are discrete day one day two and so, on they're not continuous it is called a,' Discrete Time Markov Chain', moreover since we are consisting, we are considering only discrete values, each of these random variables, is not in RN it's in 0 comma 1to n. Right? So, it's a discrete random variable. So, this is a discrete time, discrete space, Markov chain. Okay? Is that fine. Okay? Why are we interested in Markov chains? We don't know yet. Right? So, we'll get there soon, for now let us just focus on some more properties, of Marko chains. And we'll soon tie it back to our original, description of approximating the expectation and what. Okay?

Refer slide time :(09:54)



- Recall that each $X_i \in \{0,1\}^n$

X_{i-1}	X_{i-2}	T_{ab}
1	1	0.05
2	2	0.06
$\vdots$	$\vdots$	$\vdots$
l	l	0.2

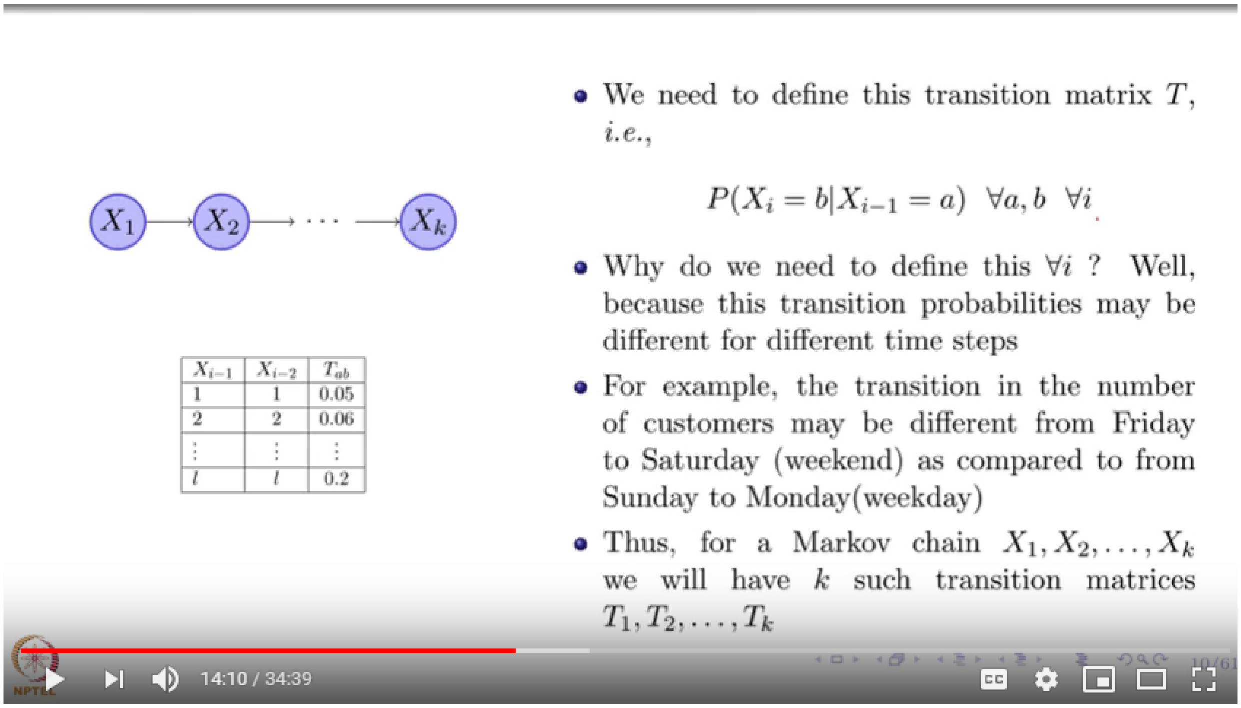
- Let us delve a bit deeper into Markov Chains and define a few more quantities
- Let us assume $2^n = l$ (*i.e.*, X_i can take l values)
- How many values do we need to specify the distribution

$$P(X_i = x_i | X_{i-1} = x_{i-1})? \quad (l^2)$$

- We can represent this as a matrix $T \in l \times l$ where the entry $T_{a,b}$ of the matrix denotes the probability of transitioning to state b from state a (*i.e.*, $P(X_i = b | X_{i-1} = a)$)
- The matrix T is called the transition matrix

So, let us delve a bit deeper into this and define a few more quantities. So, now this remember that X I, can take on to raise to n possible values, wait just I'll just call 2 raise to n as L. So, that it simplifies some things for me. So, X I can take L values, where we all agree that L is equal to 2 raise to n. Okay? Now how many values, do we need to specify the following distribution? How many values do you need to specify? This L square right I mean order L square L square minus 1 why L square? For each value of X minus I minus 1, I need to specify the probability of it being a 1or a 0. Right? So, this is how it is. Right? So, I have X I minus 1, which can take on values 1 to L. Now for each of these I need to define, what is the probability of X I minus 2 taking one of these values. Right? So, I'll have L square total values, is that fine, avian is ok with that. And we can actually represent this by a matrix, of size L cross L and we'll call this as a transition matrix, why transition matrix? Because, it tells us given a state, a at time step I minus 1, what is the probability of transitioning to state B at time step I, is that fine. So, that's why I will call this a transition matrix, it is a huge matrix and as L cross L which is raise to n cross 2 raise to n entries, is it fine. Okay? Now and the entry T a comma B in this matrix, although actually I have flattened the matrix here. But you can imagine this is a matrix, where you have the X I minus 1 values as the rows and the excise values as the column and the IG at entry tells you the probability of transitioning, from state I at time step I minus 1, to state J at time step by, everyone gets that definition Okay? So, that's a matrix and we'll call this a transition matrix, it's a huge matrix ok let's be aware of that. Okay?

Refer slide time :(11:57)



X_{i-1}	X_{i-2}	T_{ab}
1	1	0.05
2	2	0.06
⋮	⋮	⋮
l	l	0.2

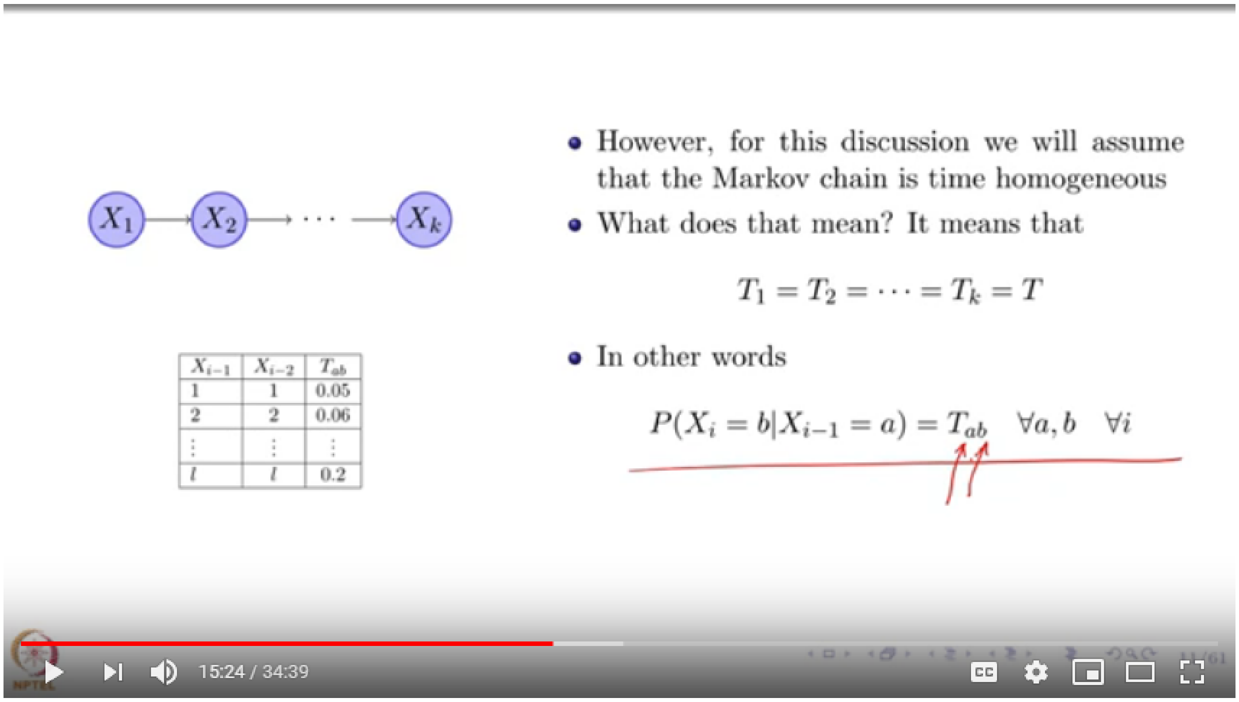
- We need to define this transition matrix T, *i.e.*,

$$P(X_i = b|X_{i-1} = a) \;\; \forall a, b \;\; \forall i$$

- Why do we need to define this $\forall i$? Well, because this transition probabilities may be different for different time steps
- For example, the transition in the number of customers may be different from Friday to Saturday (weekend) as compared to from Sunday to Monday(weekday)
- Thus, for a Markov chain $X_1, X_2, \ldots, X_k$ we will have k such transition matrices $T_1, T_2, \ldots, T_k$

Now we need to define this transition matrix, for every for all a comma B, that means, not the matrix actually, we need to define T a B, for all T comma for all a comma B that means I need to define this matrix, which has all the a comma B values in it, for all I why am I saying for all I? Say I have defined it for time step 3. Right? So, it tells me how do I transition from value a at time step 2 to, value B at time step 3. Now what does it mean to define it for time step 4, just repeat what I said how to transition from, a at time step 3 to be a time step 4. Okay? So, why do I need to do it for every time step? Why can't just have one T do you get the question how many if you get the question? Please raise your hands. Now the Markov property tells us that it depends, on the previous guy. Right? But the Markov property does not tell it that at every point, it depends in the same way on the previous gray, I'll give you a trivial example where it will become clear, think of restaurants for example, Monday, Tuesday, Wednesday you see what would happen? For example, transition properties may be different, for different days. Right? So, if you take the restaurant, the transition from Friday to Saturday. Right? When you are moving into a weekend, may be different as compared to the transition from Saturday to Monday, do you get that. So, from Friday to Saturday, the property of transitioning from- hi would be, hi but on Sunday to Monday, the transition probability of transitioning from high to high, would below. Because you would expect for your customers on a Monday, does that make sense that's why it depends on the time step, where you are in the time step? So, in general you need to define it, for every step of the chain. Right? So, you need a transition matrix, for every time step. So, you need these T 1,T 2, TK as if life was not complicated, enough 1 T was 2 raise to n cross 2 raise to n, now I'm asking you to define several such, T's. Okay? Of course we'll bring in our Savior which is some assumption.

Refer slide time :(14:14)



X_{i-1}	X_{i-2}	T_{ab}
1	1	0.05
2	2	0.06
$\vdots$	$\vdots$	$\vdots$
l	l	0.2

- However, for this discussion we will assume that the Markov chain is time homogeneous
- What does that mean? It means that

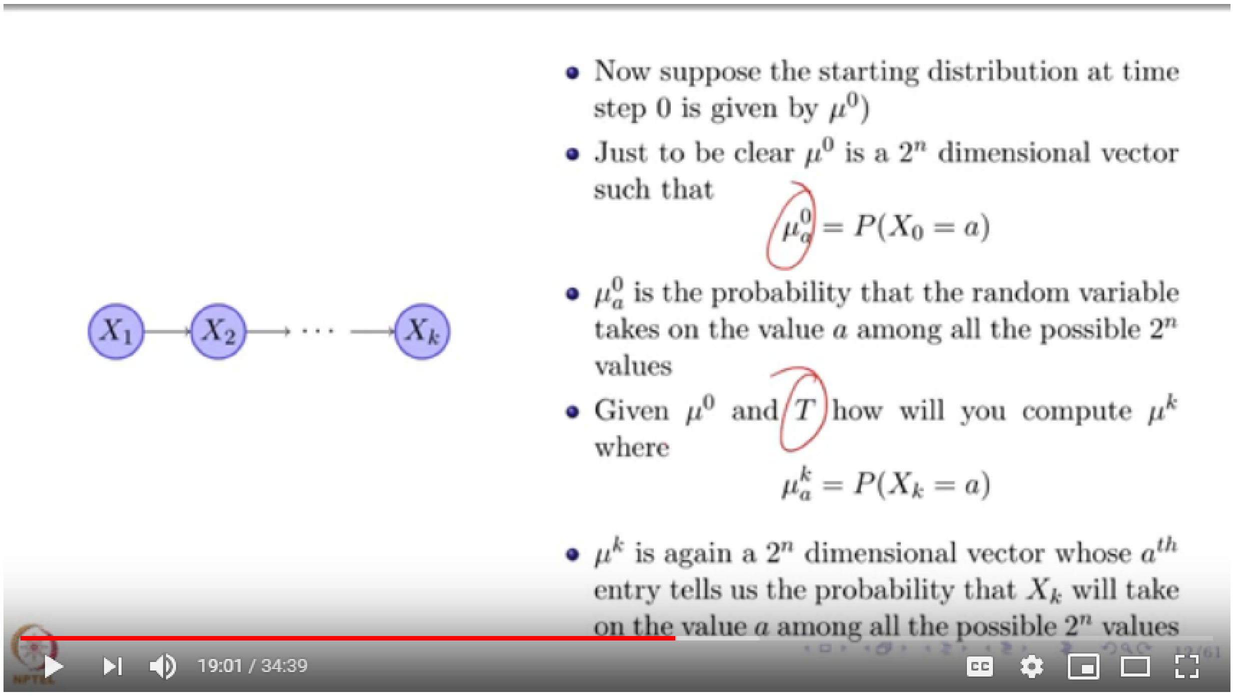
$$T_1 = T_2 = \cdots = T_k = T$$

- In other words

$$P(X_i = b | X_{i-1} = a) = T_{ab} \quad \forall a, b \quad \forall i$$

So, we'll make an assumption that and will not make an assumption, actually. So, there are some Markov chains, which are time homogeneous. So, that means that for such Marko chains, the transition matrix remains the same across time steps. And this again you could imagine, is various scenarios, we're transitioning from one state to another, remains the same every day. Right? So, if we look at on a given day and of course under certain assumptions, how many people who had taken a bus today we'll take a train tomorrow and vice versa. Right? So, you would expect over a larger population, for these transition probabilities to be more or less stable irrespective of the day, of course you can again have the weekend argument there. But still a we could have several cases in which you can assume that the Markov chain is time homogeneous. Right? And that's a simplified that just simplifies our life, because we just need to care about, one T in that case. Okay? So, that's what this homogeneous, Marko chain means that the probability of transitioning, from state a, at time step I minus one to state B at time step I, is the same irrespective of what the value of I is, is the same on Monday, Tuesday, Wednesday, Thursday and so on. Right? Okay? So, this is known as a time homogeneous Markov chain.

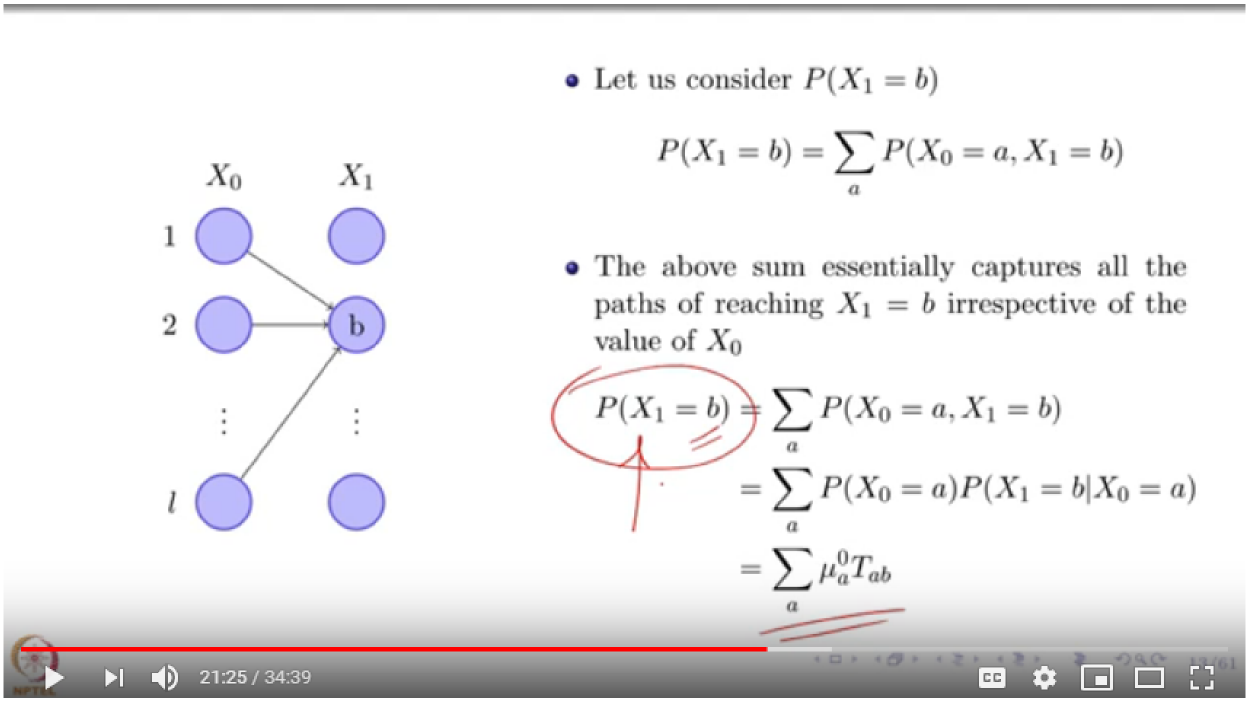
Refer slide time :(15:28)



- Now suppose the starting distribution at time step 0 is given by μ^0)
- Just to be clear μ^0 is a 2^n dimensional vector such that
$$\mu^0_a = P(X_0 = a)$$
- μ^0_a is the probability that the random variable takes on the value a among all the possible 2^n values
- Given μ^0 and T how will you compute μ^k where
$$\mu^k_a = P(X_k = a)$$
- μ^k is again a 2^n dimensional vector whose a^{th} entry tells us the probability that X_k will take on the value a among all the possible 2^n values

So, we have so, far what have we described we have described a Markov chain. We observe that it is a discrete time Markov chain; we also observed that it is a discrete time discrete space Markov chain. And now it's a discrete time, discrete space, time homogeneous, Markov chain that's what we are focusing on. Okay? Now suppose the starting distribution at time step 0, is given by mu zero. Now what does this mean? What, what do we mean by starting distribution at time step 0? What will this be a distribution over? How many values will this be a distribution over a to raise to n values. Right? So, it tells us that of all the to raise to configurations, which are possible, what's the probability of any of these configurations on day 0 at the starting time step. Right? So, how many elements does mu have. Right? So, 2 raised to n is what it has. Right? So, in particular the8th entry of this vector, tells me that the random variable X 0 that means at time step 0, it will take on the value a. Right? So, this remember that X itself is a vector. Okay? So, and there are 2 raised ton such vectors possible. So, what mu zero a tells me is that of all these 2 raised to n possible vectors, what is the probability? That X 0 will take on the value a, which is one such vector, is that clear. Okay? Let's be very clear about the dimensions of things here. So, that there is no confusion. So, the dimension of T was L cross L the dimension of MU is L. Okay? Okay? Fine now given mu 0 and transition matrix T, how will you compute mu K, where mu K is defined as the following that the 8th entry of MU K ,tells me the probability that the XK random will variable, will take on the value, you get the question what is this mu K denoting. Okay? So, let's see you have X 1, X 2 up to XK and then maybe even more. Right? Now again at X K this guy can take any of the possible, 2 raise to n values. Right? So, mu K, is a distribution over these 2 raise to n values, is that clear. So, I am asking you at time step K what is the distribution over these 2raise to n values. What, what values can the random variable X K take with certain probabilities. Right? And in particular the a at entry of mu K, again tells me, the probability that the random variable X K will take on the value a, is that clear. Okay? Now what's the dimension of MU K ,everyone, everyone L. Okay? So, it's 2 raised to n and whose a in't ND tells us the probability that I just define. Okay? Now let us consider X 1. So, our first time step was X 0. Now I am at X 1 and here's what my question, was I'll just repeat the question someone has given mu 0, someone has given me the transition matrix. And now I'm interested in MU K in general. But in specific I'll start with mu 1,I'm I want to find out, what is mu 1what's the distribution over these 2raise to n values? At time step 1that's what I'm trying to find out. Okay?

Refer slide time :(19:03)



- Let us consider $P(X_1 = b)$

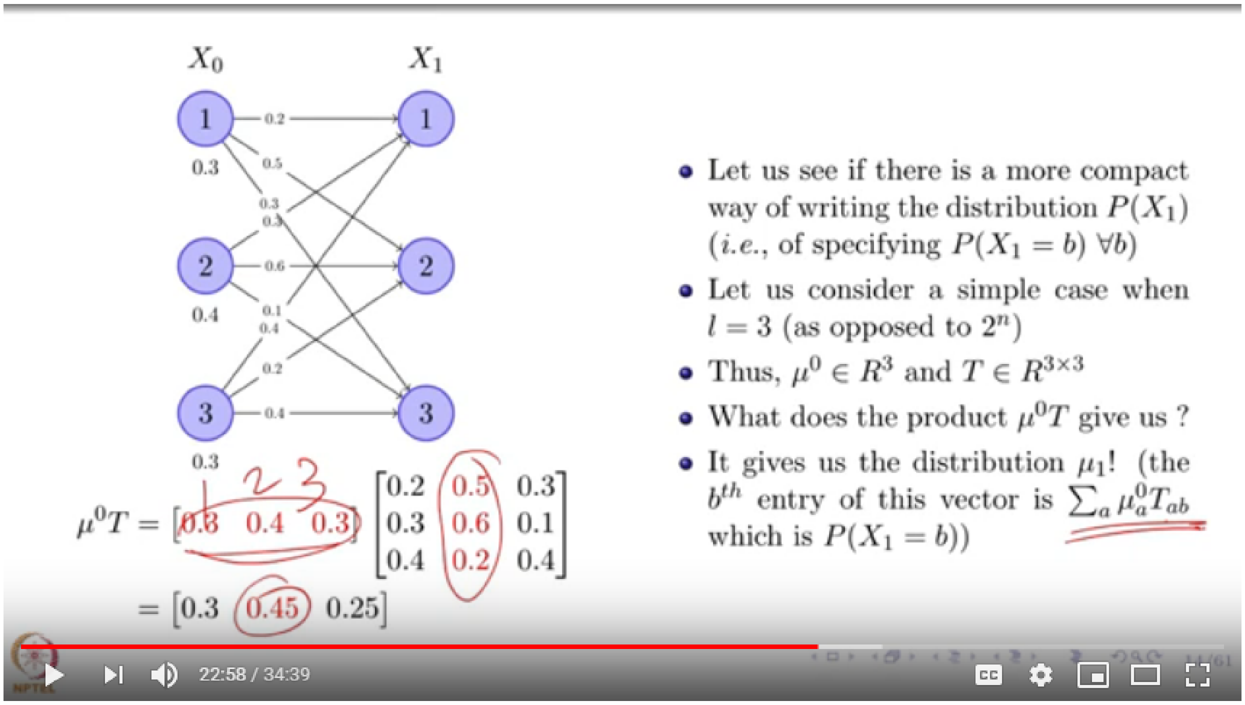
$$P(X_1 = b) = \sum_a P(X_0 = a, X_1 = b)$$

- The above sum essentially captures all the paths of reaching $X_1 = b$ irrespective of the value of X_0

$$\begin{aligned} P(X_1 = b) &= \sum_a P(X_0 = a, X_1 = b) \\ &= \sum_a P(X_0 = a) P(X_1 = b | X_0 = a) \\ &= \sum_a \mu_a^0 T_{ab} \end{aligned}$$

So, one of the values, are one of the entries, of MU K is x1 equal to B. So, which entry of MU K, is this the B at entry of, MU K. Right? Just as mu K a was the royalty that x1 can take on value, sorry not mu 1 this is the first you know is that fine. Okay? Now let us consider P ofx1 equal to B, I can write it as the following, what have I done here, introduce a variable and then marginalize Doris this is a fair operation. Okay? Fine now let's see. So, what does the above some captures, it actually captures that what is the priority of reaching, x1 equal to B, irrespective of where I start with at x0equal to a any of the x0 that I start with I'm just summing over all the possible paths, of reaching the value B at time step one, starting with any value at time step 0, is that fine, that's what this sum is capturing. Okay? And now of course this joint distribution, how will I factorize it? I can just use the simple chain rule, at this point. Right? So, X 0 equal to a, X 1equal to B, X 0 equal to a, what is this everyone? Mu 0 of a what is this everyone, everyone? T ab. So, that means this is the following quantity. Right? So, the probability, that X 1 will take on the value B, at time step 1, can be written as a neat function, of your starting distribution and your transition matrix, if I know these two I can compute it. Right? Of course as I've just shown it for time step 1, I need to convince you about this for any arbitrary time step also. Okay? So, but at least for time step 1, all of us are convinced, but we did we do not end there, let us come up with a more compact way of writing, this so, this what I have done here is tell told you how to compute, one of these probabilities. But how many such probabilities, am I interested in L of these. Right? I want to know x1 equal to ABCD up to how many other entries I have. Right?

Refer slide time :(21:30)



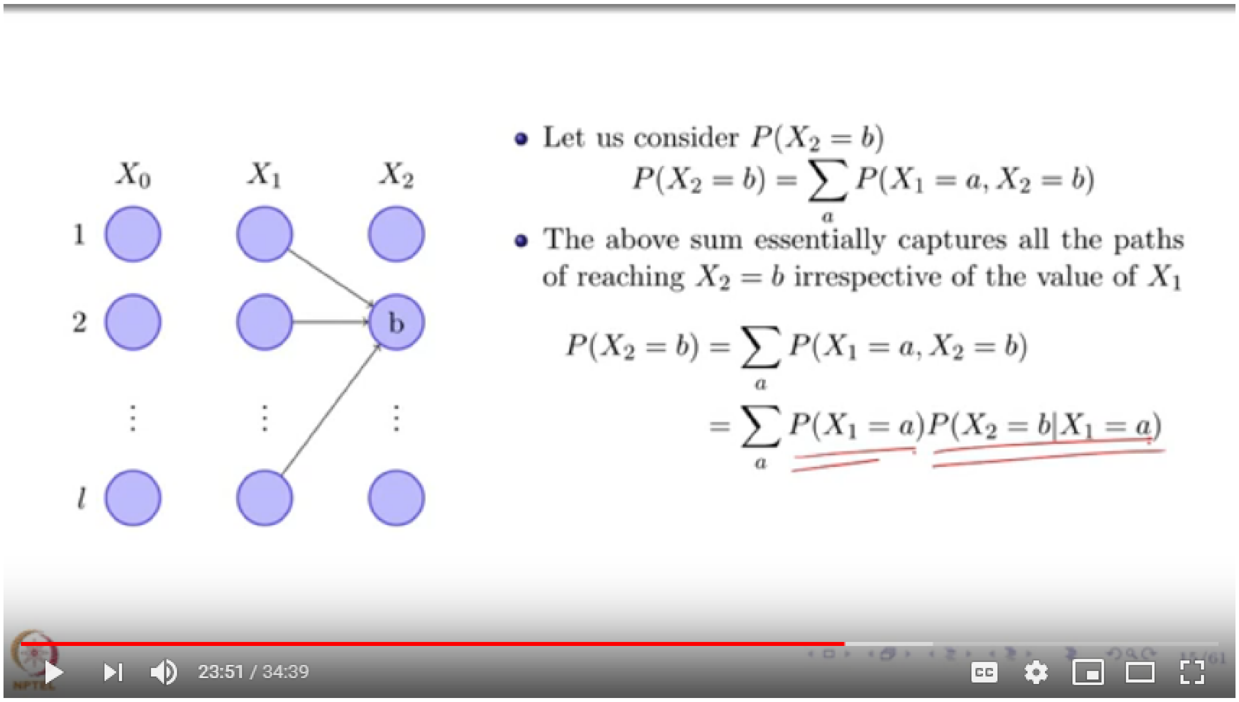
$$\mu^0 T = [\overset{1}{0.3}\ \ \overset{2}{0.4}\ \ \overset{3}{0.3}] \begin{bmatrix} 0.2 & 0.5 & 0.3 \\ 0.3 & 0.6 & 0.1 \\ 0.4 & 0.2 & 0.4 \end{bmatrix}$$

$$= [0.3\ \ 0.45\ \ 0.25]$$

- Let us see if there is a more compact way of writing the distribution $P(X_1)$ ($i.e.$, of specifying $P(X_1 = b)\ \forall b$)
- Let us consider a simple case when $l = 3$ (as opposed to 2^n)
- Thus, $\mu^0 \in R^3$ and $T \in R^{3\times3}$
- What does the product $\mu^0 T$ give us ?
- It gives us the distribution μ_1! (the b^{th} entry of this vector is $\sum_a \mu_a^0 T_{ab}$ which is $P(X_1 = b)$)

So, let us consider a simple case, where L is equal to 3 instead of 2raised to n. So, I have 3 possible states and I can transition to the same three states, at time step 1 and these edges between these states, tell you the transition probabilities. So, these are the TA B's and the values that you see under the nodes, are what mu zero. Right? So, these are mu 0 1 2 3. Okay? Sorry mu 0 of, state 1 state 2 and state 3 is that fine. Now let's see what are and of course again we should always be careful about the dimension. So, this is our cube and this is 3 cross 3 everyone is fine with that. Okay? Now what does this product actually give us if I take this is a vector and a matrix if I take its product, what will I get? What will I get everyone are you sure you'll get me one. Okay? Good. So, in fact if you look at it Right? So, if I look at the second entry of the resultant vector, first of all, all of us are sure that this is a vector by a matrix so the result would be a vector. Now if I look at the second entry of this vector, it's actually a sum of, an element by some of the multiplication of these two vectors. And that's exactly what I had here. Right? Mu zero a, where a goes from one to three, multiplied by p1 b T to be and t3 b and that's exactly what the second entry of this vector capture. Right? All of you get this how many fake clear with this. Okay? Good. So, we can write, this compactly as mu 1, is equal to MU 0 into T and I'm sure all of you see where we are headed with this, all of you see that. Okay?

Refer slide time :(23:19)



- Let us consider $P(X_2 = b)$

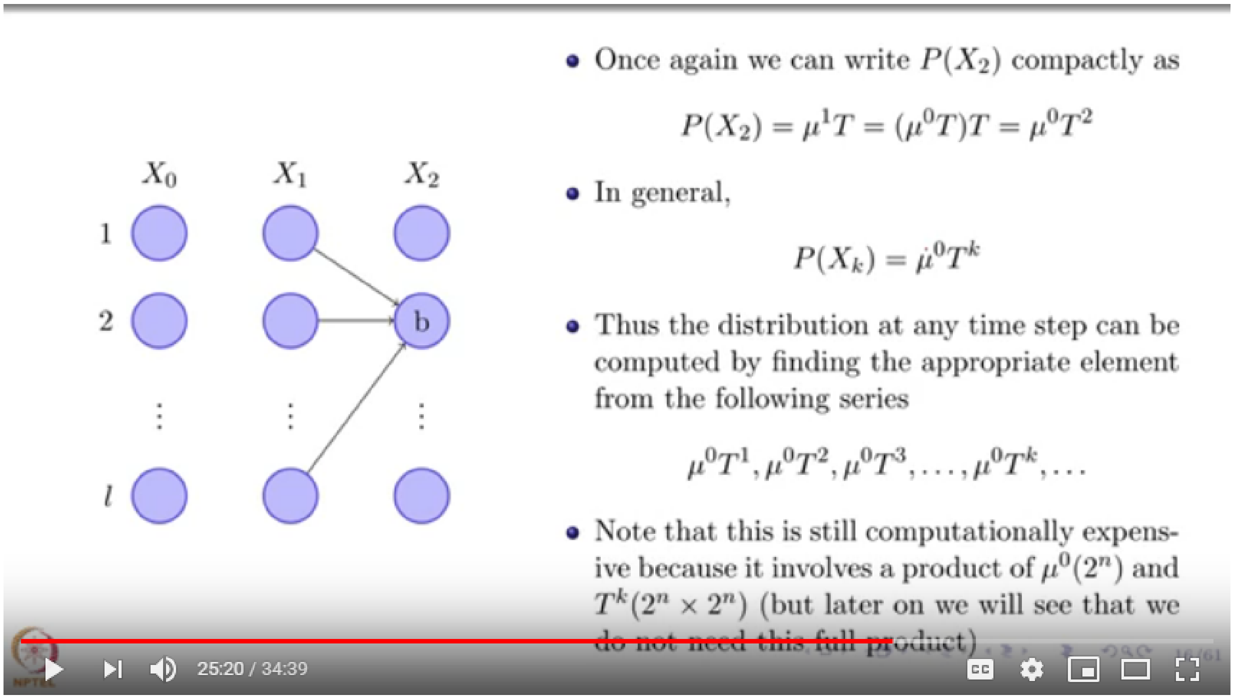
$$P(X_2 = b) = \sum_a P(X_1 = a, X_2 = b)$$

- The above sum essentially captures all the paths of reaching $X_2 = b$ irrespective of the value of X_1

$$\begin{aligned} P(X_2 = b) &= \sum_a P(X_1 = a, X_2 = b) \\ &= \sum_a \underline{P(X_1 = a)}\,\underline{P(X_2 = b|X_1 = a)} \end{aligned}$$

But we will still do it let us consider P x2equal to B, again I will follow the same but I see P, I can write it as the following Joint Distribution. So, again I have introduced the random variable X 1and then marginalized over it. Okay? Again the same story it captures, how I can reach B starting from any of the values of x2, again I will factorize it using the same, chain rule what is this? U1 of a, what is this? Ta b it should have been t1 a B, but we have assumed that t1, t2, t3 everything is T. Okay? So, this is again mu 180 TB.

Refer slide time :(23:55)



- Once again we can write $P(X_2)$ compactly as

$$P(X_2) = \mu^1 T = (\mu^0 T)T = \mu^0 T^2$$

- In general,

$$P(X_k) = \mu^0 T^k$$

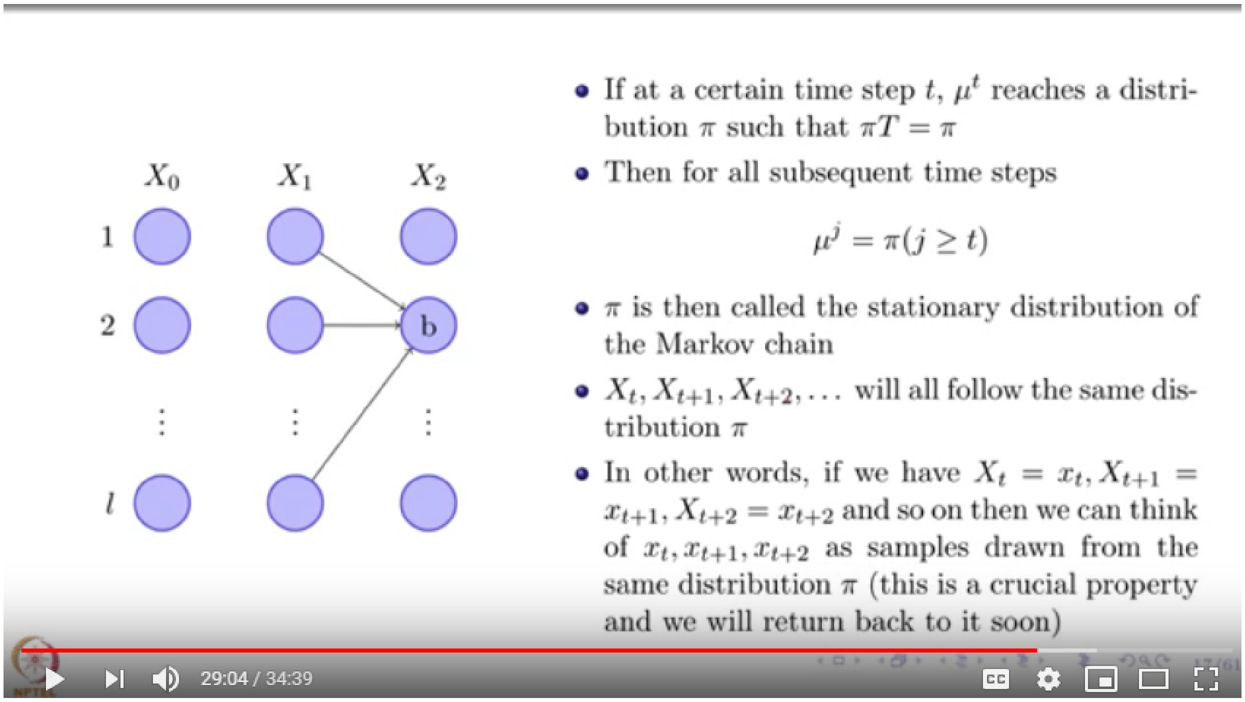
- Thus the distribution at any time step can be computed by finding the appropriate element from the following series

$$\mu^0 T^1, \mu^0 T^2, \mu^0 T^3, \ldots, \mu^0 T^k, \ldots$$

- Note that this is still computationally expensive because it involves a product of $\mu^0(2^n)$ and $T^k(2^n \times 2^n)$ (but later on we will see that we do not need this full product)

Again I can write, the entire thing compactly as mu and T. Right? What is mu 1, mu naught into T. Okay? So, in general, if I ask you mu K, this was actually mu 2, in general if I ask you mu K, what is it going to be mu naught into T raise to K. Right? So, given the initial distribution and the transition matrix, you can compute the distribution, at any given time step. Okay? For the given assumptions that is a discrete time, discrete space and time homogeneous, Markov chain. Okay? If it was ma not time homogeneous I couldn't have, used this. Right? Because the T's would have been different. Okay? Fine so, a distribution at any time step, can be computed by finding the appropriate product, from this series, music t1, T Square, t cube. And so, on and this is all very easy. Right? Does anyone see a problem with this I'm assuming that you are given mu R mu naught and TI, just need to give you these two quantities and you are set for life. What’s the problem here? What’s the problem here? Computation. Right? So, mu naught, is a 2 raise to n dimensional vector, T is a 2 n cross 2 n matrix, you need to do this expensive computation, to get these distributions. Right? But later on we will see that we don't need the full product, we can do something smartly and get by without actually computing, this product, but still get the distribution. Okay?

Refer slide time :(25:20)



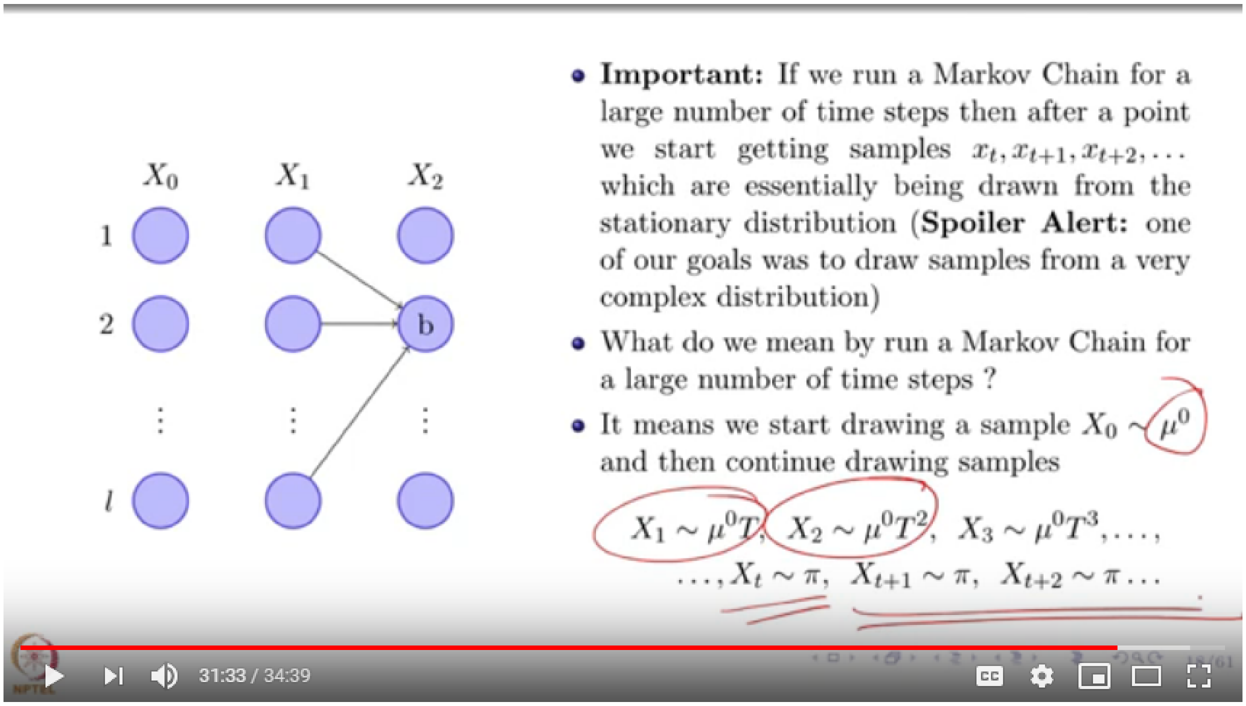
- If at a certain time step t, μ^t reaches a distribution π such that $\pi T = \pi$
- Then for all subsequent time steps

$$\mu^j = \pi (j \geq t)$$

- π is then called the stationary distribution of the Markov chain
- $X_t, X_{t+1}, X_{t+2}, \ldots$ will all follow the same distribution π
- In other words, if we have $X_t = x_t, X_{t+1} = x_{t+1}, X_{t+2} = x_{t+2}$ and so on then we can think of x_t, x_{t+1}, x_{t+2} as samples drawn from the same distribution π (this is a crucial property and we will return back to it soon)

So, that was one set of definitions, now we will just do a few more definitions and then eventually get to the point at some point, like we only have April 26, as the last degree. So, if at a certain time step T, suppose mu T, reaches a distribution PI, such that PI into T is equal to PI, what does this seemingly profion statement, Ellis I hear the word stationary over one honey many years. Okay? So, mu 0, mu 1,mu 2 and so on. Right? I was just computing all of these and I can keep computing it and see at mu K, whatever I reached, I'm just calling it by PI I can call it anything right I'm just calling it PI. Now mu K has a certain property that if I multiply it by T, I get mu K back. Okay? And here of course instead of MU K, I am just denoting it by PI. Now if this happens, what is mu k plus 1? It's mu K into T. Right? But I already know that's equal to, MU K is that fine. So, now from this time step onwards, the distribution for XK plus 1, XK plus 2, XK plus 3 and so on is going to remain the same, is that fine. Right? So, if this happens for all the subsequent time, steps you have reached the same distribution. So, let's put this in context, we had assumed that T was the same, across all the time steps. But the Meuse were not the same for all the time steps, they were different, we were computing them, as mu 0, mu 1, mu 2and so on. But now we are saying that if at a certain time step the above condition holds and from that time step on words, even the Mews become the same right so for all J greater than equal to that time step, your mu is going to be the same which is PI. And pi is then call the stationary distribution of the Markov chain. Okay? Again some definitions, I'm not saying why this would happen or will this happen and so on, under what conditions this can happen? All that we will see later. But if it this happens then this is what it would mean. Okay and now you can see that, X T, XT plus 1, XT plus 2 all these random variables, which are at time step greater than T, actually greater than equal to T, follow the same distribution, which is PI. So, instead of MU I am calling that,' Stationary Distribution', as pi. Okay? And now if I think of samples being drawn. Right? So, if I'm looking at samples from time step T. So, remember that, at time step T, the random variable X T can take all of these stories to n possible values. But it can take these values not uniformly, but according to a distribution, which is this mu T is the distribution, which is pi is a distribution. Right? Again at time step T plus 1 it can take all of these to raise to N values, but according to this distribution pi. So, now because all of these time steps have the same distribution, you can think of these samples, X T, small X T, small XT plus 1, small XT plus 2 and so on, all of them coming from the same distribution, which is pi, is that clear, does that make sense, how many of you are clear with this? I see a blank Nikita, you didn't get it let's call it T for now. So, so the key thing to note here, is once you read the stationary distribution, after that even though it seems that you are having a different random variable at every time step, you are essentially drawing from the same distribution, all these random variables come from the same distribution, all these samples come from the same distribution. Okay? That's a crucial property, I'm going to turn back, to way. Okay?

Refer slide time :(29:08)

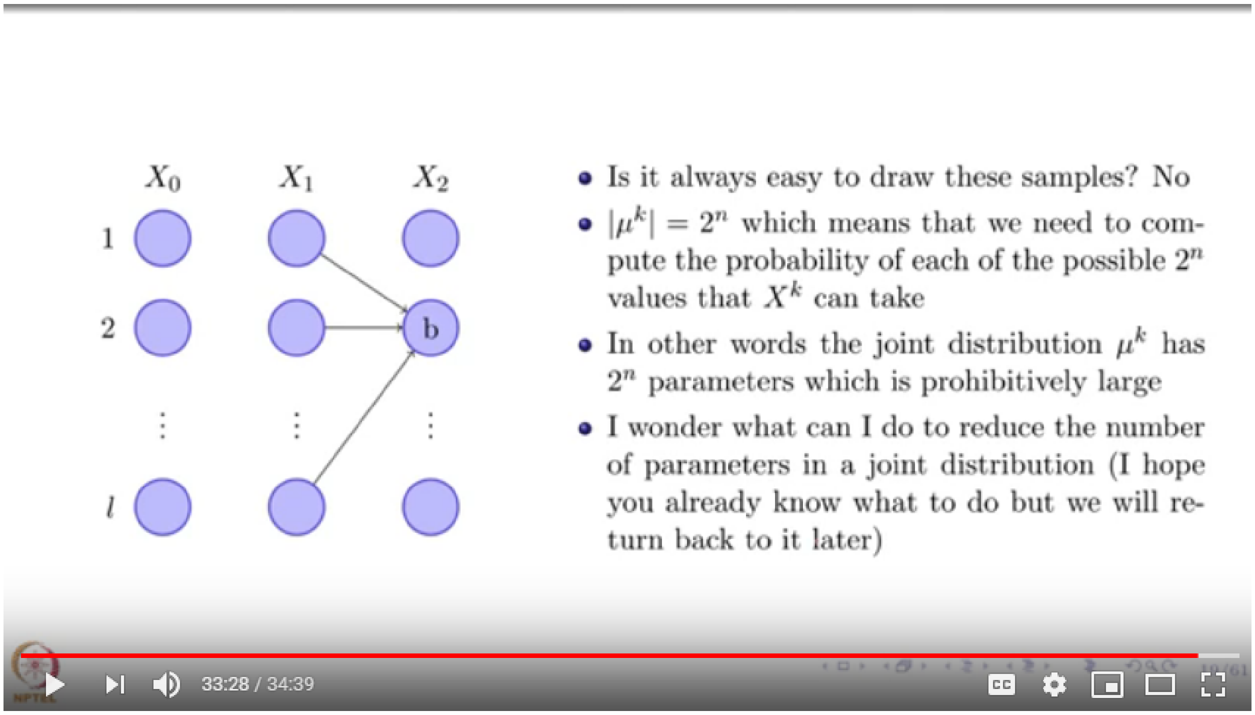


- **Important:** If we run a Markov Chain for a large number of time steps then after a point we start getting samples $x_t, x_{t+1}, x_{t+2}, \dots$ which are essentially being drawn from the stationary distribution (**Spoiler Alert:** one of our goals was to draw samples from a very complex distribution)
- What do we mean by run a Markov Chain for a large number of time steps ?
- It means we start drawing a sample $X_0 \sim \mu^0$ and then continue drawing samples

$$X_1 \sim \mu^0 T, \; X_2 \sim \mu^0 T^2, \; X_3 \sim \mu^0 T^3, \dots, \\ \dots, X_t \sim \pi, \; X_{t+1} \sim \pi, \; X_{t+2} \sim \pi \dots$$

So, now this is the important take away from here, if we run a Markov chain, for a large number of time steps, such that it reaches its stationary distribution, from that time step onwards, all these samples come from the same distribution. And why do we care about this what's the spoiler alert here, one of the goals that we had was, sample from that intractable distribution. Now whatever you have learned about Markov chains, can you give me the next part of the story right can you tell me where will head eventually, we will try to set up a Markov chain such that, such that it's stationary distribution is, is what is the required distribution? What is the distribution that we wanted to sample from, P of X. So, if you can set up a Markov chain, such that it's stationary distribution is P of X then, if you run the Markov chain for a large number of stem steps, we'll know that it will reach the stationary distribution and once that happens, whatever samples are coming, they are as if they were drawn from this P of X, even though we can actually not compute P of X. Right? Because there are a lot of ifs here and we look at all of these ifs one by one, but that's the story that we are headed towards. Right? And that's why this setup of Marko chains. Okay? And what do we mean by running a Markov step for a large number of steps, a Markov chain for a large number of time steps, what does it mean? It means that it's starting from time step zero, well you will draw a sample according to the distribution mu naught, then at time sample time step one, you will draw a sample according to this distribution, at time step two you will draw a sample according to this distribution and you will continue doing this, at some point you will of course reach the stationary distribution. And from there on if you continue drawing samples then you'll get samples from the stationary distribution PI. So, this is what I mean by running the Markov chain for a large number of time steps, keep drawing samples from the chain, at every time step according to the relevant distribution and what's that relevant distribution, it's either mu naught or mu 1 or mu 2 or mu 3and so on. I'm ignoring all the computational intractability of mu 1 mu2 and so on. But we'll try to simplify it as we go along. Right? But that's what I mean by running a Markov chain for a large number of time steps. Okay?

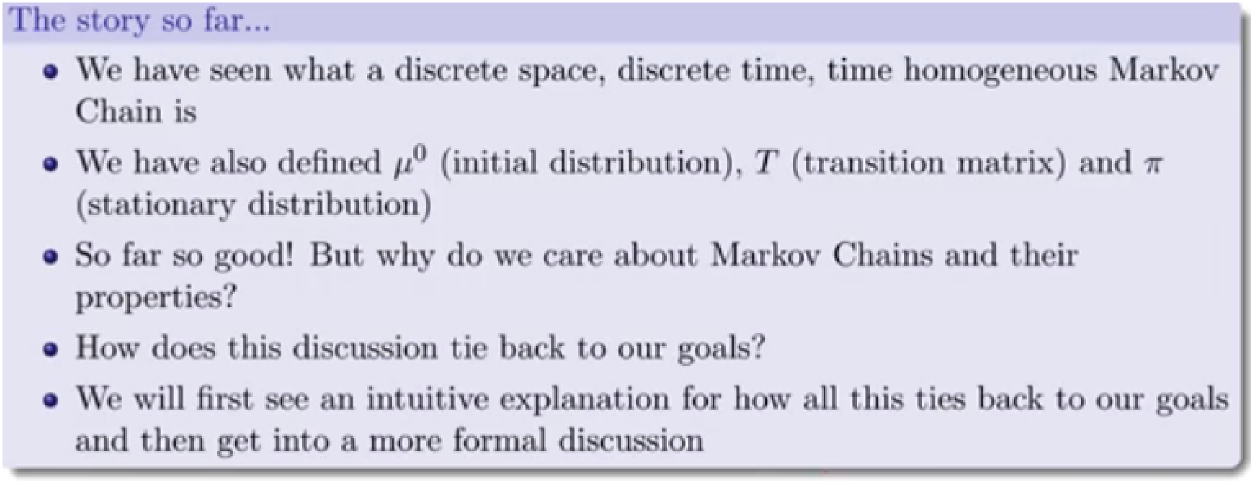
Refer slide time :(31:38)



- Is it always easy to draw these samples? No
- $|\mu^k| = 2^n$ which means that we need to compute the probability of each of the possible 2^n values that X^k can take
- In other words the joint distribution μ^k has 2^n parameters which is prohibitively large
- I wonder what can I do to reduce the number of parameters in a joint distribution (I hope you already know what to do but we will return back to it later)

Is it always easy to draw these samples so I very conveniently written that I will draw X 1 from this distribution, is it always easy to draw these samples, when in doubt always say no. Right? I guess give you an example that this mu 1requires, a 2 n cross 2 raise to n cross2 raise to n cross 2 raise to n multiplication. Right? Mu 1 is mu naught into T. So, if I want to compute the distribution mu 1 and then sample from that distribution, I have to do this, inefficient matrix multiplication. Right? So, the answer it's not always easy, but we'll have to focus on things, which are easy. So, that we reach our goal and the reason it's not easy is that mu K the dimension of MU K is 2 raised to n, which means that we need to compute all these2 raise to n values and then draw according, to that distribution which is not going to be easy. Now when you have these large number of parameters, what can you do to reduce the number of parameters? You have a distribution, which has a large number of parameters, it's a joint distribution and you want to reduce the number of parameters of the distribution, what will you do? Factorization, factorization depends on independent solutions did, we make any independence assumptions. What are we talking about. Okay? That's not a wrong answer, but bigger picture what are we talking about? RBMS. Okay? Can we make any assumptions there, independence assumptions, no we made some assumptions there. Right? So, that's a missing piece of the story that we'll get to. Okay? So, we need to somehow be able to compute these joint distributions efficiently and at some point we'll rely on this factorization that we'll have. Okay? Okay?

Refer slide time :(32:32)



The story so far...

- We have seen what a discrete space, discrete time, time homogeneous Markov Chain is
- We have also defined μ^0 (initial distribution), T (transition matrix) and π (stationary distribution)
- So far so good! But why do we care about Markov Chains and their properties?
- How does this discussion tie back to our goals?
- We will first see an intuitive explanation for how all this ties back to our goals and then get into a more formal discussion

So, that ends the first module and we'll quickly summarize, what we have done so, far we have this discrete space, discrete time, time homogeneous Markov chain, we have also defined the starting distribution, mu zero the transition matrix P and the stationary distribution, PI all this is fine, why do we care about Markov chains and their properties, this is not clear, so, far how does this description tie back to our goals, even this is not clear. So, first we will see an intuitive explanation for why all this is required. Right? And then we will get into a more formal discussion and then actually make it useful. Okay?