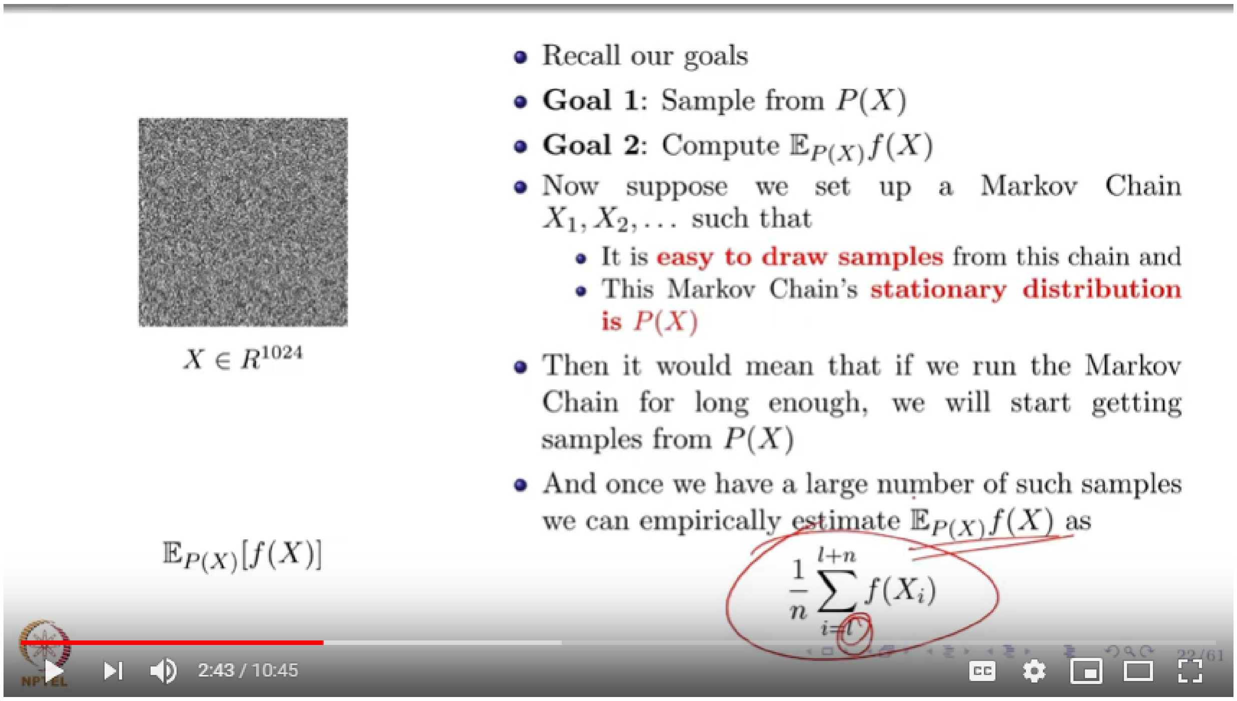
Lecture 19.2
Why de we care about Markov Chains?

So, let's start the next module, where we will talk about, why do we care about Markov chains, in the context of RBMS. So, that's what we are going to do in this module. Okay?

Refer slide time :(0:23)



$X \in R^{1024}$

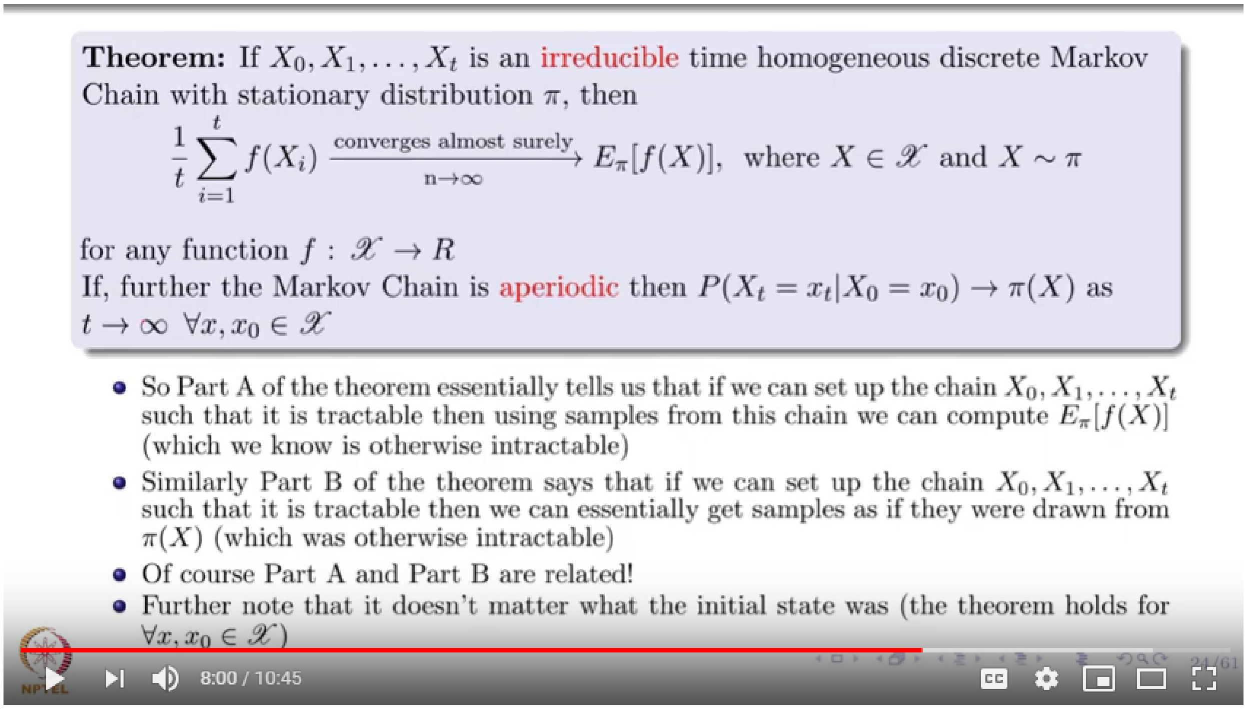
$\mathbb{E}_{P(X)}[f(X)]$

- Recall our goals
- **Goal 1**: Sample from $P(X)$
- **Goal 2**: Compute $\mathbb{E}_{P(X)} f(X)$
- Now suppose we set up a Markov Chain $X_1, X_2, \ldots$ such that
  - It is **easy to draw samples** from this chain and
  - This Markov Chain's **stationary distribution is** $P(X)$
- Then it would mean that if we run the Markov Chain for long enough, we will start getting samples from $P(X)$
- And once we have a large number of such samples we can empirically estimate $\mathbb{E}_{P(X)} f(X)$ as

$$\frac{1}{n}\sum_{i=l}^{l+n} f(X_i)$$

So, recall our goals goal one was sample from P of X goals two was compute this interact, able expectation and of course both the goals are related. Okay? Now suppose we set up a Markov chain, X 1 X 2 up to whatever such that what is the condition, the dash of this- is equal to dash stationary distribution, of this Markov chain, is P of X. Right? Okay? And further it is easy to draw samples from this chain. Right? There's no point in computing, in constructing a chain such that at each point if I want to sample something from the chain, is as hard as sampling from the original distribution. And that's what we saw in the setup. Right? Because at every time step I had to compute this mu 1 mu 2 and so on which was as hard as anything else right because, I have to do this very expensive matrix multiplication. Right? So, I should be able to set up with some Markov chain, such that it is easy to draw samples from that chain. And the stationary distribution of this chain, is the distribution that I care about, P of X and now because it's easy to draw from this chain, once I reach PI and once I start drawing samples from there, it would be easy for me to draw samples from the distribution that I care about, does that statement make sense, irrespective of whether it's clear, how to do it or not at least the goal makes sense. How many for clear with what I just said? Please raise your hands, I in above. Okay? So, then it would mean, if these conditions hold that if it is easy to sample from the chain and if the stationary distribution is P of X then if you run, this chain for a large number of time steps, then eventually we'll start getting samples from P of X, is it fine. Okay? And once we have that once, I reach a time step L, at which how somehow know, that this is a stationary distribution, then from that point I onwards I can take n samples, because I know these n samples would come from the distribution, that I care about which is P of X. So, then I can approximate this expectation, by this empirical expectation. Right? So, that’s what I'm interested, in doing I am interesting setting a chain, such that a stationary distribution, is the distribution that I care about, then run this chain for long enough. So, that I get samples from this chain and then use those samples, to empirically compute the expectation that I want, is that here. Okay?

Refer slide time :(02:49)



Theorem: If $X_0, X_1, \ldots, X_t$ is an irreducible time homogeneous discrete Markov Chain with stationary distribution π, then

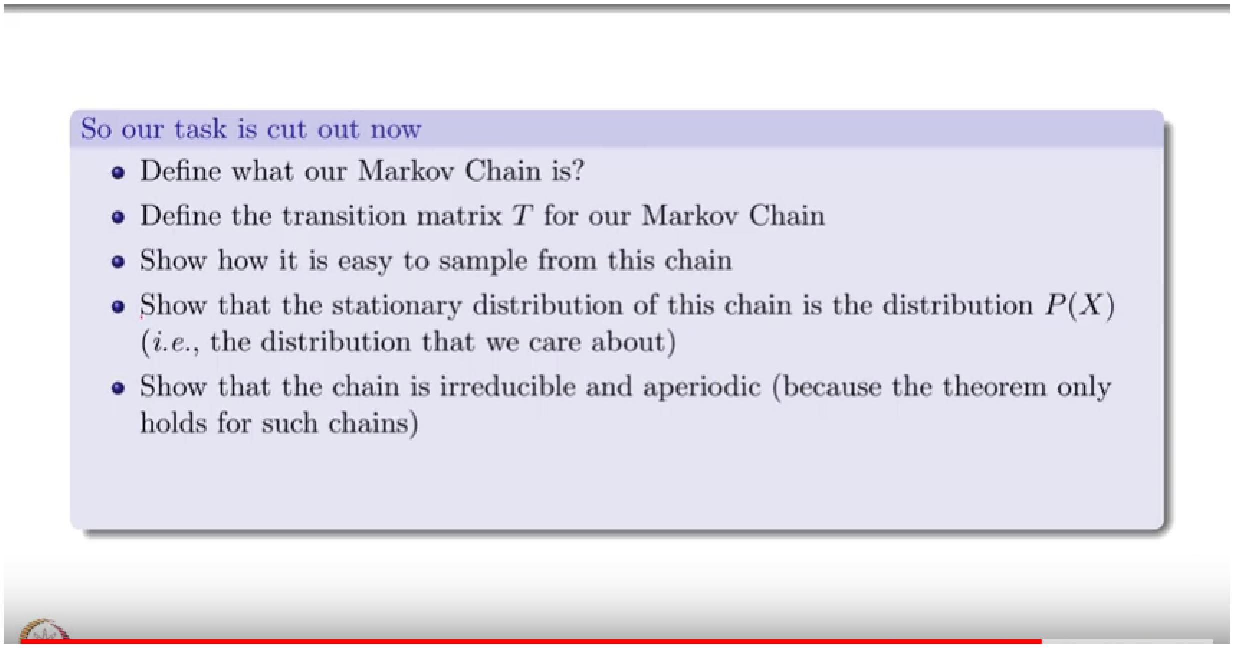
$$\frac{1}{t}\sum_{i=1}^{t} f(X_i) \xrightarrow[\text{n}\to\infty]{\text{converges almost surely}} E_\pi[f(X)], \text{ where } X \in \mathscr{X} \text{ and } X \sim \pi$$

for any function $f : \mathscr{X} \to R$
If, further the Markov Chain is aperiodic then $P(X_t = x_t|X_0 = x_0) \to \pi(X)$ as $t \to \infty \ \forall x, x_0 \in \mathscr{X}$

- So Part A of the theorem essentially tells us that if we can set up the chain $X_0, X_1, \ldots, X_t$ such that it is tractable then using samples from this chain we can compute $E_\pi[f(X)]$ (which we know is otherwise intractable)
- Similarly Part B of the theorem says that if we can set up the chain $X_0, X_1, \ldots, X_t$ such that it is tractable then we can essentially get samples as if they were drawn from $\pi(X)$ (which was otherwise intractable)
- Of course Part A and Part B are related!
- Further note that it doesn't matter what the initial state was (the theorem holds for $\forall x, x_0 \in \mathscr{X}$)

Fine so, now we will get into a more formal discussion and my formula mean I will bring in some theorems. Right? So, if X naught, X 1, up to X T, is an irreducible, time homogeneous discrete when I say discrete I mean discrete time as well as discrete space, Markov chain with stationary distribution pi. So, far nothing really groundbreaking you know all this, this is a this is the Marko chain that we just saw, then this theorem, tells us that, if I take samples from this chain. Okay? Then the expectation of a function f of X, under the stationary distribution PI, I can get it by just doing this empirical estimate. So, this empirical estimate, will almost surely converge to the true expectations, remember this is all asymptotically, that means as n tends to infinity right and we are never going to do n tends to infinity, but still is the statement of the theorem clear and if I have such a chain and if the stationary distribution, of the chain is PI, then if I take a large number of samples from this chain. And empirically compute the expectation. Right? This is the same if you remember; this quantity is the same as how I had estimated the average weight of the expected weight of the population. Right? I had taken some T samples and I just taken the average of the weight that's exactly I had done it empirically. So, if you have samples from the distribution, I will empirically estimate the expectation and I can be almost sure, that is going to converge to the true expectation, if n tends to infinity, for all X capital X belonging, to what is this actually? The stylish X what is it in R case? Sample space what is it in R case? The answer is simple, 0 comma 1 raise to M. Right? But I'm just writing in and because the sample space could be anyone, anything, right it's not that it's only related to binary sample spaces it could be anything. So, for us at0 1, raise to n but this could be any sample space. And X belongs to the stationary distribution R X is X follows the stationary distribution pi and this holds for any function, which maps from the sample space to a real value. Right? So, in R case the sample space is 0 to 1 raise to n. So, any function which takes me from this sample space to R, the above statement holds true for R. Right? In particular, it will hold true for whatever we had inside those two nasty-looking expectations, is that fine, is it clear. So, remember the origin of all this is those two expectations that we care about, when we compute the gradient of the log-likelihood, with respect to W IJ, we had these two expectations. And expectations are nothing but e with respect to a distribution, of some function. So, irrespective of what that function? Is the above theorem will hold true. Okay? Further if the chain is a periodic, then the probability of XT, taking on the value small XT, given some value of x 0, approaches PI of X, what does this mean? So, irrespective of where you started from, at time step T you are interested in finding on, what is the probability? That the random variable capital XT, will take on some value, let's say small XT that is the same as so, P of XT, equal to XT, irrespective of where you started. Because it's for all small X T's and for all small X not, the same as PI of X that means, is the same as P of XT equal to XT. Right? Where P this, P I'll just call it P one is a stationary distribution. Right? So, I can run the chain for a long time and after a point, I can be sure that even though the samples are coming from this distribution, they are actually coming from, my stationary distribution, that’s just a fancy way of saying the same thing that, once you read the stationary distribution, the samples are coming from, the stationary distribution .Everyone is clear with, this. Okay? That’s why as this again asymptoticle. Right? So, as n tends to infinity, so that means it is you start from the starting state, some point it is the stationary distribution and now n tends to infinity rate. So, now at that point, everything starts coming from this stationary distribution. So, Part A of the theorem essentially tells us that, if we set up the chain, such that it's stationary distribution is the distribution that we care about, then we have a clean empirical way, of approximating the expectation, that we care about, part two of the theorem, which is if further, tells us that if you set up, the Markov chains, as that is stationary distribution is the distribution that we care about, then after some point we'll start getting samples from, this distribution of course Part A and Part B are related, because we can approximate, the expectation, because the samples that we are getting are from the two distribution that we care about. And it does not matter, where you start from because, the theorem holds for all small X naught and small XT belonging to your sample space. Okay? Is that fine.

Refer slide time :(08:05)



So our task is cut out now

- Define what our Markov Chain is?
- Define the transition matrix T for our Markov Chain
- Show how it is easy to sample from this chain
- Show that the stationary distribution of this chain is the distribution $P(X)$ ($i.e.$, the distribution that we care about)
- Show that the chain is irreducible and aperiodic (because the theorem only holds for such chains)

So, now given this setup, our task is cutoff. We first need to decide, what our Markov chain is going to be. Okay? I'll tell you, what the Markov chain is going to be, to define a Markov chain, I should tell you what the transition matrix is, because a Markov chain depends on the transition matrix. Okay? I need to tell that it is I need to show, that it is easy to sample from this chain, I do not need to do this expensive computations, every element, of the chain, can come efficiently, without any expensive computations, I need to show you that the stationary distribution of whatever chain I compute, is going to be P of X That's the distribution that we care about. Okay? I need to show you that the chain is irreducible and a periodic, why was those are the two things which I had not defined. And the theorem relies on those two property. Right? The chain has to be irreducible; it has to be a periodic I have not defined what these mean yet? I'll define it soon and I also give you an intuition, for why whatever chain we set up is A periodic an irreducible. So, if I show you all of this, then we are done. Right? Then I can get you samples from the distribution that you care about. And then you can compute the expectation that you care about. Okay? And for ease of notation, this capital X that I have been talking about, which was this random variable. I am going to use X to denote, the random variables that we had in the case of RBMS, which were these M visible units and N hidden variables all of this collectively, I'm going to call as X. And so, X actually belongs to 0 comma 1 raise to, n plus M and I will refer to the individual elements of X, as X 1 X 2. So, these are not the elements of the chain, this are the dimensions of one particular random variable. Right? So, the random variable itself is of size n plus m. And these are the elements of that random variable. Is this change of notation clear to everyone? Otherwise you not understand anything going forward. Okay? So, the Club sum total of the visible and the hidden variables, I am denoting it by capital X it contains, all the hidden and the visible variables. Is that clear? Okay?