

Artificial Intelligence
Prof. Mausam
Department of Computer Science and Engineering
Indian Institute of Technology - Delhi



Lecture-60
Bayesian Networks: MCMC with Gibbs Sampling

(Refer Slide Time: 00:25)

MCMC with Gibbs Sampling

- Fix the values of observed variables
- Set the values of all non-observed variables randomly
- Perform a random walk through the space of complete variable assignments. On each move:
 1. Pick a variable X
 2. Calculate $\Pr(X=\text{true} \mid \text{all other variables})$
 3. Set X to true with that probability
- Repeat many times. Frequency with which any variable X is true is its posterior probability.
- Converges to true posterior when frequencies stabilize significantly
 - stationary distribution, mixing

Markov chain needs to be 1. aperiodic, 2. irreducible



Now we are going to forget whatever we have learnt and just learn the MCMC algorithm. MCMC algorithm stands on a Markov chain Monte Carlo. Monte Carlo is a term that is used for sampling based algorithm. And Markov chain you should already know you must have studied in a probability class. In a Markov chain there is a state and then the probability of transitioning to the new state which depends only on the previous state if it is a first order Markov chain in that what we will be studying.

So we create a chain of states and then do whatever we want to do and we ask the question what is a stationary distribution of such a Markov chain. And all of the state theory gets relevant here. I am not going to teach out all that theory. If you want to understand this algorithm deeply go home read upon Markov chain read upon how to prove stationary distribution check this algorithm again and understand this algorithm and then try to prove that the stationary

distribution of what I am just teaching you is going to be the posterior probability distribution of the Bayesian network or read the book.

But let us learn the algorithm together that our goal for today what is going to do is a square of fix the observed that the beauty I am never considering anything for the first step I have fixed all the evidence variable. Everything that is going to happen is going to happen with the evidence variable set the values at a given to us. Now I am going to randomly generate a state. So I am going to set values of all non observed variables randomly.

Just completely randomly too much about it does not matter. Now I am going to perform a random walk a Markov chain random walk through the space of complete variable assignment through the specific complete states. And in each move I am going to pick of variable X and I am going to sample that variable again. So all the $n - 1$ variables will remain the same I am going to only change one variable. And that one variable will be calculated re-flip based on probability of X equal to true given all the other variables.

And set to be true that probability and toss a coin and set X to be true. And I repeat many, many times and the frequency with which any variable X is true is its posterior probability. This is the algorithm that is it. We will spend 2 minutes, 3 minutes understanding this. But that is it is nothing more to this of course you do not understand this yet deeply enough but algorithm itself it should be clear by this now nothing more that is going on.

And again we fix all the evidence variables will never worry about the evidence variables again. They always set to the values that a given to us. So, probability of B given C , C is true, neither will I pick C nor will I sample C . C will be true throughout my process. I will set all other variables randomly, so, randomly set burglary, randomly set randomly earth quack, randomly set alarm and randomly set newscast that my starting state.

Now, I will pick a random variable, let sat pick A and I am going to sample A again given all the other variables including C equal to true. So, C will always be true in the given part it is never going to change it will always be sampling based on the evidence, but not just the evidence but

all the variable. And think about why do we sample with all the variables think about it you will get the answer. And we sample it and we set A to be true or false with that probability and repeat, repeat means pick another variable let say burglary.

Again sample burglary true with all the other variables and then set burglary to this value and we keep going. And believe it or not this convergence to the true posterior distribution when frequency stops changing significantly. However because we have started from a specific start state initially my sample should be highly biased. So, I will get rid of the initial number of 100000 samples or order longhand samples or something like this.

And there are theoretical properties on how many samples it takes to get to this Markov chain getting mixed. Mixed means irrespective of which state I started from I have now come to more or less to the same distribution, that is the right distribution I want to get to that is called the mixing time. It is a theoretical concepts, in practice we say we are going to get rid of first 100000 sample that is called the burn in time.

Burn in is a practical measure for mixing time with the theoretical phenomena. The stationary distribution means the distribution when the probabilities stop changing. We have reached the point where each sample will happen with the same probability that is called the stationary distribution. And now you have to prove stationary distribution exist. Number one you will reach it. Again under what conditions will you reach it, again I am sweeping something under the rack.

And so for example Markov chain needs to be non periodic, aperiodic and I forget that 2 properties that are may aghartic or something and 3. When you reach the stationary distribution, what is the distribution you are going to converge to and you can prove that will be the posterior distribution. All of you, who are slightly more mathematically inclined I really encourage you to read more about this. This is the most important algorithms in the 2000's era.

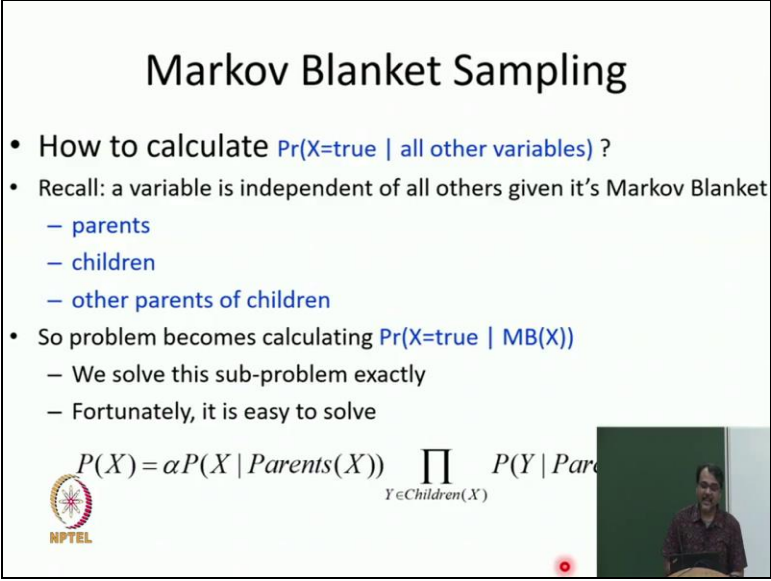
This is called Gibbs sampling by the way. You would see Gibbs sampling used left, right and c centre again and again. And then its extension something called metropolis-hasting or importance sampling all of these algorithm are still very important. Whenever I have to samples

from some distribution these algorithms become extremely, extremely important. So, please read more about it.

Now one thing how about 0.2 here calculate probability of X equal to true given all the variables? Is it easy to do this? How do I even calculate probability X of a given all the variable is it not that an exponential process? Can somebody see how would I compute this? Most of them are independent, what is your name? Jay; Jay says most of them are independent which rule is Jay using when he says that most of them are independent? Markov blanket.

So given the Markov blanket everything else is independent. This is given everything including Markov blanket so only the variables that are in the Markov language are important for this particular sampling style that is the beauty of this.

(Refer Slide Time: 07:48)



Markov Blanket Sampling

- How to calculate $\Pr(X=\text{true} \mid \text{all other variables})$?
- Recall: a variable is independent of all others given it's Markov Blanket
 - parents
 - children
 - other parents of children
- So problem becomes calculating $\Pr(X=\text{true} \mid \text{MB}(X))$
 - We solve this sub-problem exactly
 - Fortunately, it is easy to solve

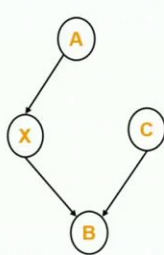
$$P(X) = \alpha P(X \mid \text{Parents}(X)) \prod_{Y \in \text{Children}(X)} P(Y \mid \text{Parents}(Y))$$

The slide includes an NPTEL logo in the bottom left and a small video inset of a speaker in the bottom right.

To compute probability of x given away all the variables I will take parents, children and other parents and children of x and not worry about anything else. Probability of x given everything is equal to probability of x given Markov blankets and that you can compute using the basic rule.

(Refer Slide Time: 08:04)

Example






$$P(X) = \alpha P(X | \text{Parents}(X)) \prod_{Y \in \text{Children}(X)} P(Y | \text{Parents}(Y))$$

$$P(X | A, B, C) = \frac{P(X, A, B, C)}{P(A, B, C)}$$

$$= \frac{P(A)P(X | A)P(C)P(B | X, C)}{P(A, B, C)}$$

$$= \left[\frac{P(A)P(C)}{P(A, B, C)} \right] P(X | A)P(B | X, C)$$

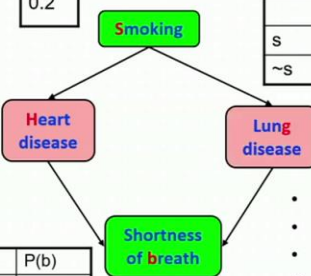
$$= \alpha P(X | A)P(B | X, C)$$

Ok so let us quickly take an example, you can read the specific computations of Markov blanket yourself it is not very complicated.

(Refer Slide Time: 08:11)

Example



P(s)	
s	0.2

P(h)	
s	0.6
~s	0.1

P(g)	
s	0.8
~s	0.1

H	G	P(b)
h	g	0.9
h	~g	0.8
~h	g	0.7
~h	~g	0.1

- Evidence: s, b
- Randomly set: h, g
- Sample H using $P(H | s, g, b)$


Suppose result is ~h

Sample G using $P(G | s, \sim h, b)$

⇒ Suppose result is g

Sample G using $P(G | s, \sim h, b)$

⇒ Suppose result is ~g


45

Let us take simple example, so if you some probably have heart disease if you probably have lung disease and if you have heart disease, you probably have shortness of breath if you have lung with shortness of breath. And let us say I give you that I smoke and I have shortness of breath. This is my evidence. And now I want to figure out what is the probability I have heart disease. So now I first start by randomly sampling the other variable.

So, I let say while it is a randomly set heart disease to be true and lung disease to be true. And now I randomly pick a variable. So let us see I randomly pick H ok. And I then use probability of H given s, g, b all the other 3 variables. And I figured out what is that probability and then I sample. Now heart disease what is the Markov blanket of heart disease? All the 3 by parents smoke, my children shortness of breathe my other parents and children and lung disease.

So then I will take all the 3 and sample H again so I compute the probability and sample H. So let us say my sample is not H. So, now my new state is s, b not h g that is my new state. I have done one step of the random walk. Now let say repeat by saying sample another variable let say sample lung. This is now when I am sampling lung disease, I will sample it given s, b which are evidence and not h because that is my previous data.


And again, let us say I sample suppose the result is g. So my new sample is s b not t h and g and that is my new sample. And I keep generating these samples and so on and so forth.

(Refer Slide Time: 10:20)

Gibbs MCMC Summary

$$P(X|E) = \frac{\text{number of samples with } X=x}{\text{total number of samples}}$$

- **Advantages:**
 - No samples are discarded
 - No problem with samples of low weight
 - Can be implemented very efficiently
 - 10K samples @ second
- **Disadvantages:**
 - Can get stuck if relationship between two variables is *deterministic*
 - Many variations have been devised to make MCMC more robust


46

So in this way, I generate lots of samples and now you have to estimate probability of X given E I will just check in how many samples express a specific value given the total number of samples and that is my answer. The beauty of the Gibbs sampling algorithm is that no samples are discarded just like likelihood waiting. But there is no problem with no wait because I am always

sampling given the evidence. I am not sampling from the prior distribution I am sampling some the posterior distribution.

So you will see the stationary distribution of Markov chain is the posterior distributions and I am actually sampling from the posterior distribution. It can be implemented extremely efficiently as you can see the algorithm was just this much so it was very, very easy implement, you can get like 10 years ago you had gotten 10000 samples a second so extremely fast. Now there is a problem with the algorithm that is very difficult to see. I will quickly tell you let say I have two variables X, Y such that Y is equal to X but X is can be true or false a probability 0.5.

So, X is uniform 0.5 and Y is exactly equal to X , so if X is true and Y is true with probability 1 and X is false, Y is false with probability 1. But when I do Gibbs sampling let us say initially both of them were 0, 0. So, when I sample Y given X I will remain at 0, 0 and when sample X given Y I will remain at 0, 0. So, I will only generate 000000 samples. And so what is going to happen I will believe that marginal probability of X is 0 which is not true it is 0.5.

I can never jumps from 0,0 to 1,1 if I start from 1,1 I will remain at 1,1. So these regions disconnected regions are called modes of the Bayesian Network. And if Bayesian network has modes when will the I have modes if my path is stuck when will be my path is cut if some probability is 0 or some probability is 1. So, it can get stuck if the relationship between 2 variables is deterministic. And lots of algorithms have been studied to make MCMC more robust to this deterministic setting.

In fact even if I have near determinism, like probability is very low even then you have to practical modes like it very difficult to move some jumps from this node this node. As long as that does not happen the MCMC is a great algorithm and work really, really well and super fast.

(Refer Slide Time: 12:58)

Other inference methods

- Exact inference
 - Junction tree
- Approximate inference
 - Belief Propagation
 - Variational Methods
 - Metropolis-Hastings



There are also other algorithms which we did not cover and if you get to ParaSingla so anybody probabilistic graphical models class you will learn about the junction tree algorithm, you will learn about belief propagation. You will learn about variational inference methods and you will also learn about extensions MCMC like Metropolis-Hasting and importance, but it is in separate class this is where we stop the discussion on inference.

In the next class will talk about learning Bayesian network parameters and structural and that will be the last class in Bayesian network after that we will move on to decision making in probabilistic environment.