01.112 Machine Learning, Spring 2018
Lecture Notes for Week 11

19. Bayesian Networks (II)

Last update: Tuesday 3rd April, 2018 15:41

Bayesian networks: graph and independence

We have already emphasized that the graph structure in Bayesian networks represents useful qualitative properties about the variables. Specifically, the graph encodes independence statements about how the variables relate to each other. We will need a criterion for reading such independence properties from the graph without consulting the underlying probability distribution. The subtlety here is that we cannot just pick any criterion we like. The probability distribution we will associate with the graph must be consistent with all the properties we can derive from the graph. Otherwise the graph would "lie" and wouldn't be useful to us.

We had previously introduced the graph as specifying the parents of each node $i = 1, \ldots, d$. In the graph, we call nodes j which start directed edges or arcs to node i as parents of i. Formally, we say that $j \in pa_i$. The choice of these parents, *i.e.*, the choice of parent sets pa_i, is unconstrained except for the fact that we cannot introduce directed cycles. In other words, the graph must be a *directed acyclic graph* (DAG). In terms of variables $X_1, \ldots, X_d$, we motivated the graph as specifying which other variables each X_i directly depends on, *i.e.*, variables whose values we would have to know prior to drawing a value for X_i. We write the set of parents as variables using notation $X_{pa_i} = \{X_j\}_{j \in pa_i}$. Once we know the parents, we can write (factor) the probability distribution over all the variables as

$$P(X_1 = x_1, \ldots, X_d = x_d) = \prod_{i=1}^{d} P(X_i = x_i | \mathbf{X}_{pa_i} = \mathbf{x}_{pa_i}) \tag{1}$$

where, for some nodes, $pa_i = \{\}$ (the empty set) and $P(X_i = x_i | \mathbf{X}_{pa_i} = \mathbf{x}_{pa_i})$ reduces to $P(X_i = x_i)$ (no other variable need to be consulted prior to sampling a value for X_i). You should convince yourself that any directed acyclic graph must have at least one node without any parents.

The above factorization resembles an application of the chain rule. First, we write the variables in some order such that the parents of each variable always come before the variable itself in the ordering. This is always possible for a directed acyclic graph (in fact, there are often a large number of such "consistent" orderings). Then we simply "drop" dependences in the conditional probabilities to get the above factorization. These simplifications represent independence assumptions about the variables. To simplify the notation, let's assume that the simple lexicographic ordering works for our graph. In other words, assume that $pa_i \subset \{1, \ldots, i-1\}, i = 1, \ldots, d$ Then, by applying the chain rule, we could always write (without any assumptions)

$$P(X_1 = x_1, \ldots, X_d = x_d) = \prod_{i=1}^{d} P(X_i = x_i | X_{i-1} = x_{i-1}, \ldots, X_1 = x_1) \tag{2}$$

When we construct the distribution for a particular graph, we are assuming that

$$P(X_i = x_i | X_{i-1} = x_{i-1}, \ldots, X_1 = x_1) = P(X_i = x_i | \mathbf{X}_{pa_i} = \mathbf{x}_{pa_i}) \tag{3}$$

where $pa_i \subset \{1, \ldots, i-1\}$. This is an independence assumption. Specifically, define $npa_i = \{1, \ldots, i-1\} \backslash pa_i$ as the "preceding non-parents". By factoring the joint distribution according to Eq. (1), *i.e.*, dropping dependences except for the parents, we make the assumption that X_i is independent of $\mathbf{X}_{npa_i}$ given values for the parents $\mathbf{X}_{pa_i}$. These independence statements, one statement per variable, would suffice to specify the directed graph (dropping all arcs from preceding non-parents). But there are many other independence statements that are implied by these. How do we read all of these off the graph directly?

Independence from the graph: D-separation

As an example, consider a slightly extended version of the alarm model we had discussed before. The model is given in Figure 1, with an additional binary variable L. This could be whether we "leave work" as a result of hearing/learning about the alarm. We will now define a procedure for answering questions such as: are R and B independent of each other? Are R and B independent of each other if we know (given) A? And so on.

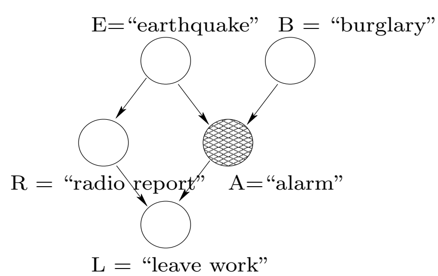
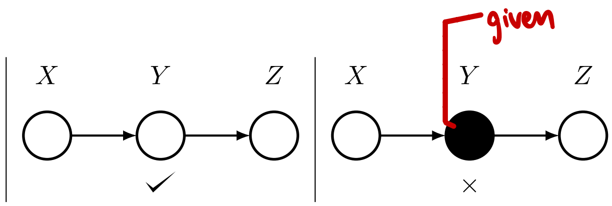


Figure 1: An example Bayesian network.

Let us first look at some simpler cases. We consider three variables X, Y and Z, and we are interested in finding the dependence/independence information between X and Z under different conditions (whether Y is given/observed or not). Three basic conditions are considered, where all three variables are connected to one another in some way.

Chain

Are X and Z dependent?



From the graph, we have:

$$\begin{aligned} P(X,Y,Z) &= P(X)P(Y|X)P(Z|Y) && (4) \\ P(X,Y) &= P(X)P(Y|X) && (5) \end{aligned}$$

Now let us look at the following conditional probability:

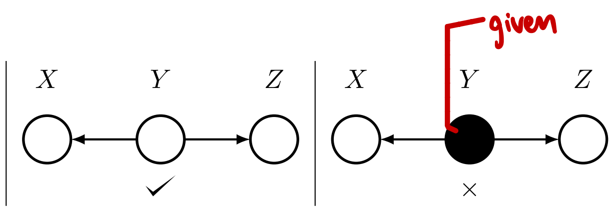
$$P(Z|X,Y) = \frac{P(X,Y,Z)}{P(X,Y)} = \frac{\cancel{P(X)}\cancel{P(Y|X)}P(Z|Y)}{\cancel{P(X)}\cancel{P(Y|X)}} = P(Z|Y) \qquad (6)$$

· Independent

This means Z and X are *independent* given Y. On the other hand, without knowing Y, the variables Z and X are *dependent*.

Common Cause

Are X and Z dependent?



The joint probability for these three variables is defined as:

$$\begin{aligned} P(X,Y,Z) &= P(Y)P(X|Y)P(Z|Y) && (7) \\ P(X,Y) &= P(Y)P(X|Y) && (8) \end{aligned}$$

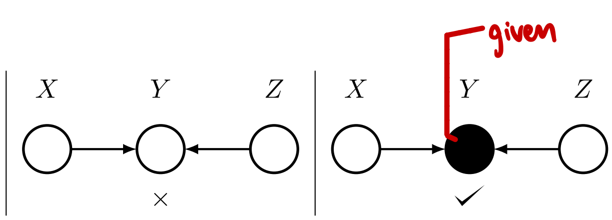
Now we look at the following conditional probability:

$$P(Z|X,Y) = \frac{P(X,Y,Z)}{P(X,Y)} = \frac{\cancel{P(Y)}\cancel{P(X|Y)}P(Z|Y)}{\cancel{P(Y)}\cancel{P(X|Y)}} = P(Z|Y) \qquad (9)$$

This means Z and X are *independent* given Y. On the other hand, without knowing Y, the variables Z and X are *dependent*.

Explaining Away

Are X and Z dependent?



The joint probability for these three variables is defined as:

$$P(X,Y,Z) = P(X)P(Z)P(Y|X,Z) \qquad (10)$$

Discussions Can we prove that $P(Z|X,Y) = P(Z|Y)$?

Okay, now let us instead look at the following *joint* probability:

$$P(X,Z) = \sum_Y P(X)P(Z)P(Y|X,Z) = P(X)P(Z)\sum_Y P(Y|X,Z) = P(X)P(Z) \quad (11)$$

This means Z and X are *independent* without knowing Y. On the other hand, Z and X are *dependent* given Y, as we have seen in the previous class when we discussed explaining away.

The notation of *d-separation* (or *directed-separation*) refers to the connectivities of two distinct sets of variables (given another set of variables) in a Bayesian network. We have so far seen the above basic cases, and next let us discuss a more general algorithm for determining *d-separation* between variables for a general Bayesian network.

Bayes' Ball Algorithm

The above independence/dependence information can be read off from the graphs directly when there are three variables. It can be shown that such independence/dependence information can be read off from a Bayesian network that involves more variables. The resulting algorithm is called the *Bayes' ball algorithm*, and it relies on the above simpler facts as building blocks.

We would like to find whether $\mathbf{x}_A$ and $\mathbf{x}_B$ are independent given $\mathbf{x}_C$, where $\mathbf{x}_A, \mathbf{x}_B$ and $\mathbf{x}_C$ are disjoint sets of random variables in a given Bayesian network.

We need to find for each variable in $\mathbf{x}_A$ and each variable in $\mathbf{x}_B$ whether they are all conditionally independent given $\mathbf{x}_C$. If so, then we conclude that $\mathbf{x}_A$ and $\mathbf{x}_B$ are conditionally independent given $\mathbf{x}_C$.

To check whether two variables X and Z are conditionally independent, we can use depth-first search to find all the paths from X to Z. Each path consists of the above-mentioned basic patterns. Imaging there is a ball (Bayes' ball) that is rolling on the path. We check if any of the path is open (*i.e.*, the path consists of patterns with a ✓ symbol below) so that the ball can reach Z from X by following that path. If there exists a path that is open, we say the two variables X and Z are *dependent*. Otherwise they are *independent*.

Markov Blanket

Assume we are interested in finding the following conditional probability:

$$p(X|\mathbf{V}_{-X}) \quad (12)$$

where $\mathbf{V}_{-X}$ is the set of variables in the Bayesian network, excluding the variable X.

Using what we have learned today, we can see that conditioning on the set of all other variables except X is equivalent to conditioning on a few variables surrounding the variable X only. This set of nodes is called the Markov blanket of the variable X.

Discussions What are the nodes that the Markov blanket consists of?

Answer: the Markov blanket of the variable X consists of X's __children__, X's __parent__ as well as the parents of X's __spouse__.

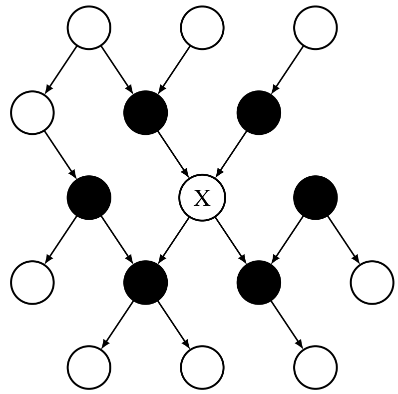


Figure 2: The conditional probability of X given all other variables is equivalent to $P(X|\mathbf{m}(X))$, where $\mathbf{m}(X)$ is the Markov blanket of X.

This observation allows us to derive the conditional probabilities easily. In fact this is a crucial observation that is useful for the Gibbs sampling algorithm, an important *approximate inference* algorithm used for inference in general Bayesian networks.

(**Optional**) The Gibbs sampling algorithm is used for generating samples $\mathbf{y} = \langle y_1, y_2, \ldots, y_n \rangle$ from a distribution given some evidence $P(\mathbf{y}|\mathbf{x})$, where $\mathbf{x}$ and $\mathbf{y}$ are sets of variables. The procedure is as follows:

1. Randomly initialize $\mathbf{y}^{(0)} = \left\langle y_1^{(0)}, y_2^{(0)}, \ldots, y_n^{(0)} \right\rangle$.
2. For $t = 1, \ldots, T$ do
   (a) For $k = 1, \ldots, n$ do
   $y_k^{(t)} \sim P(y_k | y_1^{(t)}, \ldots, y_{k-1}^{(t)}, y_{k+1}^{(t-1)}, y_{k+2}^{(t-1)}, \ldots, y_n^{(t-1)}, \mathbf{x})$ (This conditional probability can be simplified using Markov blanket.)
   (b) Collect the t-th sample as $\left\langle y_1^{(t)}, y_2^{(t)}, \ldots, y_n^{(t)} \right\rangle$
3. Return the collection of samples.

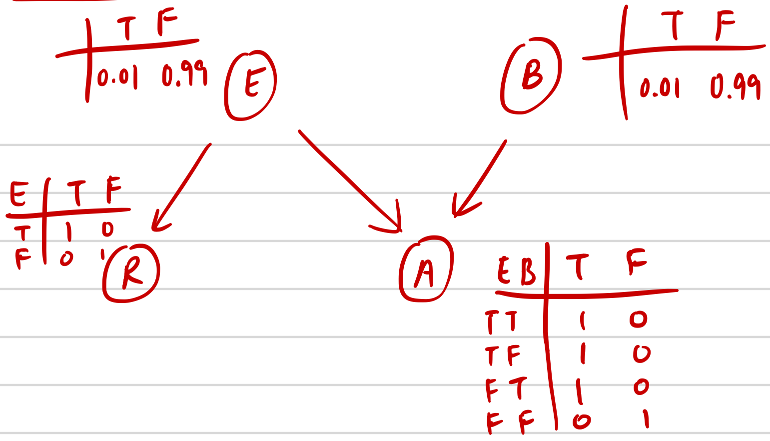
This algorithm can be used to perform approximate inference in Bayesian networks.

Learning Objectives

You need to know:

1. How does the Bayesian network capture dependence information between different variables using a DAG representation.
2. How to read off the dependence and independence information between different variables given evidences from the DAG directly by using the Bayes' ball algorithm.
3. What is the Markov blanket of a variable in a Bayesian network.

W11 01 Recap (Directed Acyclic Graph)



	T	F
E	0.01	0.99

	T	F
B	0.01	0.99

E	T	F
T	1	0
F	0	1

EB	T	F
TT	1	0
TF	1	0
FT	1	0
FF	0	1

$$P(X_1, \dots, X_n) = \prod_{i=1}^{n} P(X_i \mid Pa(X_i))$$

$P(E=e, R=r, B=b, A=a)$
$= P(E=e) \cdot P(B=b) \cdot P(R=r \mid E=e, B=b) \cdot P(A=a \mid E=e, B=b)$

Side Track to Inference Problem

1. $P(B=T \mid A=T) > 0.5$

where $P(B=b, A=T)$
$= \sum_{r,e} P(B=b, A=T, R=r, E=e)$

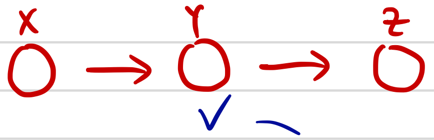
2. $P(B=T \mid A=T, E=T)$
$= \dfrac{P(B=T, A=T, E=T)}{P(A=T, E=T)}$

$= \dfrac{0.01 \times 0.01}{0.01} = 0.01 \neq 0.5$, $P(B=T)$ and $P(E=T)$ are dependent ie not independent

Conclusion: if solid line seen, 2 heads not independent
if alarm not heard, leave work independent of burglar alarm.
→ Check usg Today Theory.

W11 02 Bayesian Network (II)

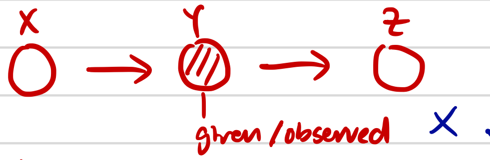
Eg 1 Chain



HW: Prove X and Z dependent.
↓
usg only 1 case.

↑

Gates.



– Independent

X ←

Hypothesis
X and Z independent given Y.
$P(X, Z \mid Y) = P(X \mid Y) P(Z \mid Y)$

LHS $= \dfrac{P(X, Y, Z)}{P(Y)}$, $P(X,Y,Z) = P(X) P(Y \mid X) P(Z \mid Y)$

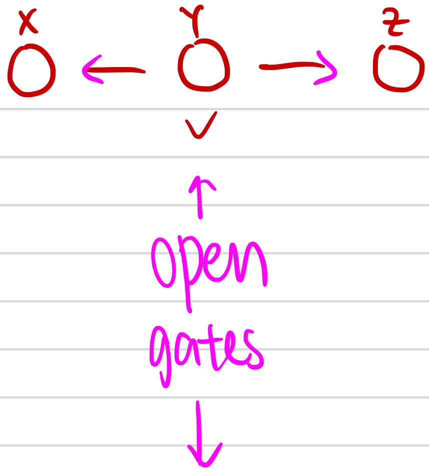
$= \dfrac{P(X) P(Y \mid X)}{P(Y)} \cdot P(Z \mid Y)$ → Read up Prob, Stats RV.

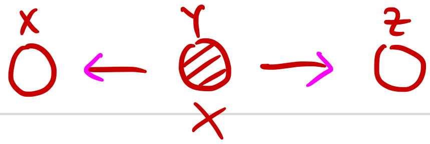
$= \dfrac{P(X, Y)}{P(Y)} \cdot P(Z \mid Y)$

$= P(X \mid Y) \cdot P(Z \mid Y)$

→ Independent

Eg2. Common Cause





Hypothesis

X and Z independent given Y.

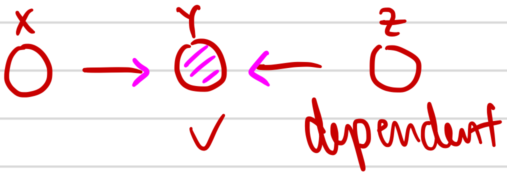
$$P(X,Z\mid Y) = P(X\mid Y)P(Z\mid Y)$$

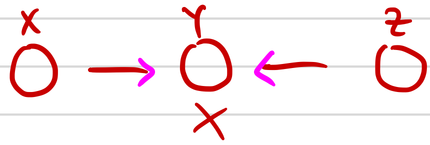
$$LHS = \frac{P(X,Y,Z)}{P(Y)}, \quad P(X,Y,Z) = P(Y)P(X\mid Y)P(Z\mid Y)$$

$$= \frac{\cancel{P(Y)}P(X\mid Y)}{\cancel{P(Y)}} \cdot P(Z\mid Y)$$

→ Read up Prob, Stats RV.

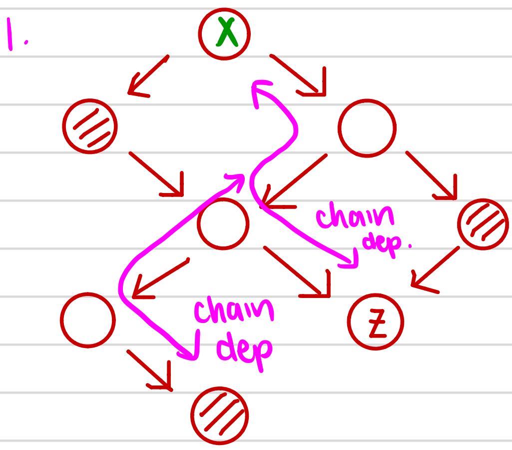
Eg3. Explaing Away



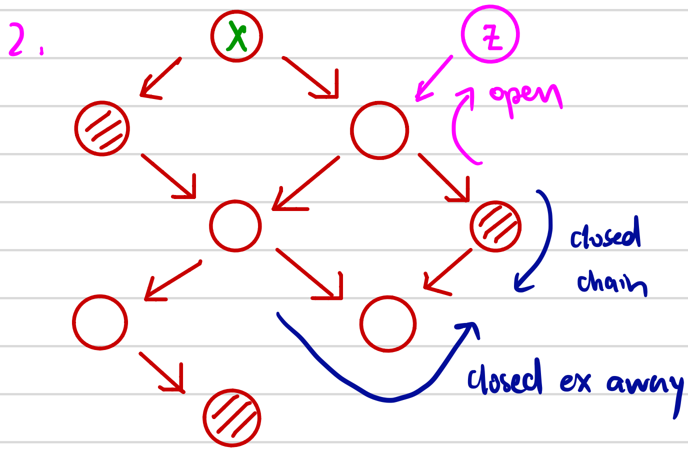


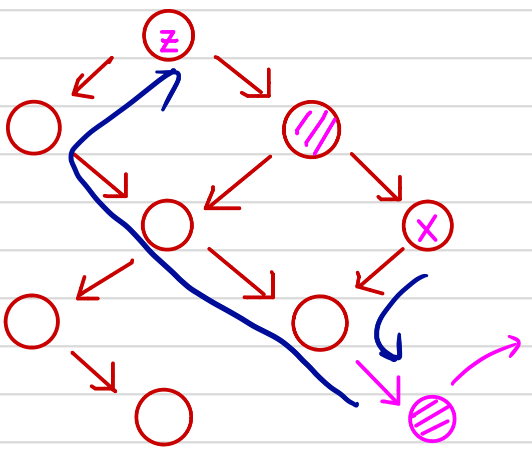
Full case Eg

1.



2.

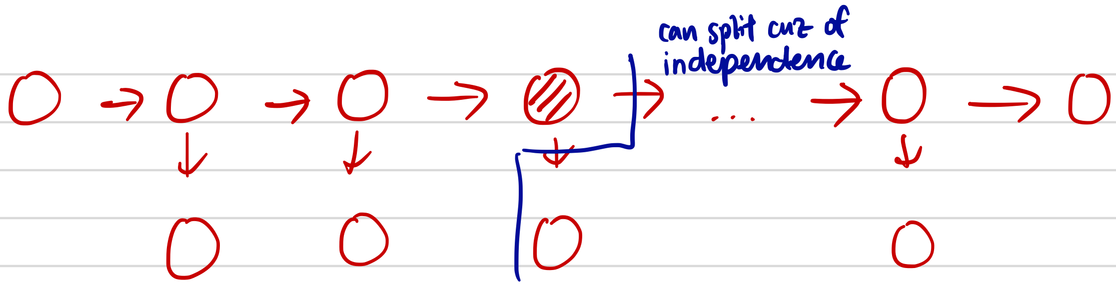




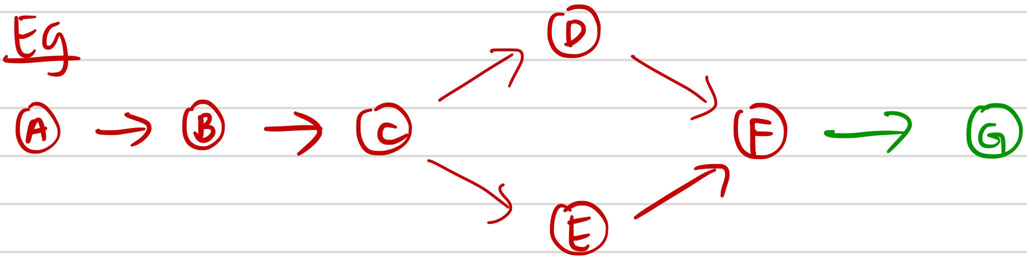
Introducing this makes independence of X and Z to dependence given blue path

BAYES BALL ALGO

HMM (Bayesian Network)



Eg



$P(B=T \mid F=T) = 0.2$

$P(B=T \mid F=T, G=T) = 0.2$

↳ because F=T given /obs.

↳ Independent case.

$P(B=T)$

$= \sum_{acdef} P(A=a, B=T, C=c, D=d, E=e, F=f)$

$= \sum_{acdef} P(A=a) \cdot P(B=T \mid A=a) \cdot P(C=c \mid B=b) \cdot P(D=d \mid C=c)\, P(E=e \mid C=c)$
$P(F=f \mid D=d, E=e)$

$= \sum_{acde} P(A=a) \cdot P(B=T \mid A=a) \cdot P(C=c \mid B=b) \cdot P(D=d \mid C=c)\, P(E=e \mid C=c)$
$\boxed{\sum_f P(F=f \mid D=d, E=e) = 1}$

$= \sum_{ac} P(A=a) \cdot P(B=T \mid A=a) \cdot P(C=c \mid B=b) \cdot \cancelto{1}{\sum_d P(D=d \mid C=c)}\ \cancelto{1}{\sum_e P(E=e \mid C=c)}$

$= \sum_a P(A=a) \cdot P(B=T \mid A=a) \cdot \cancelto{1}{\sum_c P(C=c \mid B=b)} \cdot$

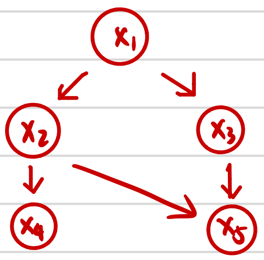
$= \sum_a P(A=a) \cdot P(B=T \mid A=a) \cdot$

variable enumeration

NP-Hard where P ≠ NP

↓ Decision Problem

↓ #P - Deterministic Counting Machine



Find $P(X_3 = T, X_2 = F \mid X_4 = T, X_1 = F)$

Step 1 : Find many samples.

Step 2 : Compute count of $(X_1 = F, X_2 = F, X_3 = T, X_4 = T)$ –(2)

Step 3 : Compute count of $(X_1 = F, X_4 = T)$ –(3)

Step 4 : Ans $= \frac{\text{Eqn}(2)}{\text{Eqn}(3)}$

General Steps for Gibbs Sampling

① Randomly Initialize $y^{(0)} = \langle y_1^{(0)}, y_2^{(0)}, \dots, y_n^{(0)} \rangle$

② For $t = 1, \dots, T$ do :

a) For $k = 1, \dots, n$ do :

$$y_k^{(t)} \sim P(y_k \mid y_1^{(t)}, \dots, y_{k-1}^{(t)}, y_{k+1}^{(t-1)}, y_{k+2}^{(t-1)}, \dots, y_n^{(t-1)}, x)$$

b) Collect the t-th sample as $\langle y_1^{(t)}, y_2^{(t)}, \dots, y_n^{(t)} \rangle$

③ Return the samples.

Sample

X_1	X_2	X_3	X_4	X_5
F	F	F	T	T
F	x	y	T	

gen x_3

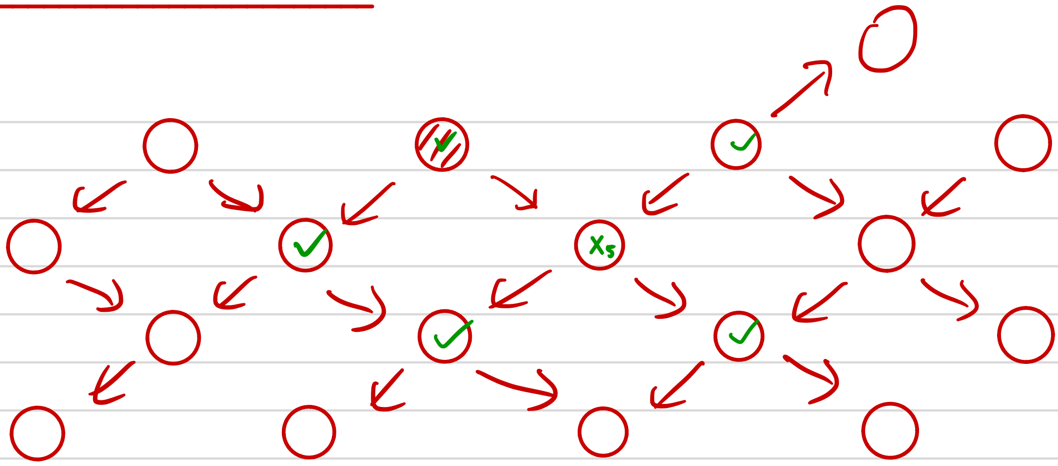
generate x_2

$X_2 \sim P(X_2 \mid X_1 = F, X_3 = F, X_4 = T, X_5 = T) \rightarrow$ maybe get T

$X_3 \sim P(X_3 \mid X_1 = F, X_2 = T, X_4 = T, X_5 = T) \rightarrow$ maybe get T

$X_5 \sim P(X_5 \mid X_1 = F, X_2 = T, X_3 = T, X_4 = T) \rightarrow$ maybe get F

Complicated Case



Markov Blanket $\leftrightarrow$ Children, Parent, Spouse.

$P(X_5 \mid V_{-X_5})$

Next Class
① Learng given Network
② Learng not given Network.