Uncertainty:

Till now, we have learned knowledge representation using first-order logic and propositional logic with certainty, which means we were sure about the predicates. With this knowledge representation, we might write A→B, which means if A is true then B is true, but consider a situation where we are not sure about whether A is true or not then we cannot express this statement, this situation is called uncertainty.

So to represent uncertain knowledge, where we are not sure about the predicates, we need uncertain reasoning or probabilistic reasoning.

Causes of uncertainty:

Following are some leading causes of uncertainty to occur in the real world.

1. Information occurred from unreliable sources.
2. Experimental Errors
3. Equipment fault
4. Temperature variation
5. Climate change.

Probabilistic reasoning:

Probabilistic reasoning is a way of knowledge representation where we apply the concept of probability to indicate the uncertainty in knowledge. In probabilistic reasoning, we combine probability theory with logic to handle the uncertainty.

We use probability in probabilistic reasoning because it provides a way to handle the uncertainty that is the result of someone's laziness and ignorance.

In the real world, there are lots of scenarios, where the certainty of something is not confirmed, such as "It will rain today," "behaviour of someone for some situations," "A match between two teams or two players." These are probable sentences for which we can assume that it will happen but not sure about it, so here we use probabilistic reasoning.

Need of probabilistic reasoning in AI:

- When there are unpredictable outcomes.
- When specifications or possibilities of predicates becomes too large to handle.
- When an unknown error occurs during an experiment.

In probabilistic reasoning, there are two ways to solve problems with uncertain knowledge:

- **Bayes' rule**
- **Bayesian Statistics**

As probabilistic reasoning uses probability and related terms, so before understanding probabilistic reasoning, let's understand some common terms:

Probability: Probability can be defined as a chance that an uncertain event will occur. It is the numerical measure of the likelihood that an event will occur. The value of probability always remains between 0 and 1 that represent ideal uncertainties.

1. $0 \leq P(A) \leq 1$, where P(A) is the probability of an event A.
1. $P(A) = 0$, indicates total uncertainty in an event A.
1. $P(A) = 1$, indicates total certainty in an event A.

We can find the probability of an uncertain event by using the below formula.

$$\textbf{Probability of occurrence} = \frac{\text{Number of desired outcomes}}{\text{Total number of outcomes}}$$

- $P(\neg A)$ = probability of a not happening event.
- $P(\neg A) + P(A) = 1$.

Event: Each possible outcome of a variable is called an event.

Sample space: The collection of all possible events is called sample space.

Random variables: Random variables are used to represent the events and objects in the real world.

Prior probability: The prior probability of an event is probability computed before observing new information.

Posterior Probability: The probability that is calculated after all evidence or information has taken into account. It is a combination of prior probability and new information.

Conditional probability:

Conditional probability is a probability of occurring an event when another event has already happened.

Let's suppose, we want to calculate the event A when event B has already occurred, "the probability of A under the conditions of B", it can be written as:

$$P(A|B) = \frac{P(A \wedge B)}{P(B)}$$

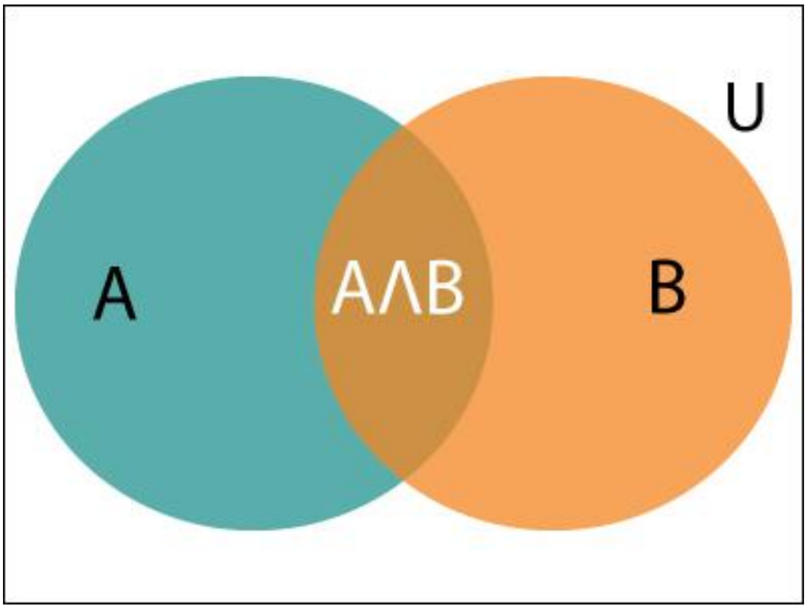
Where P(*A*∧*B*)= Joint probability of a and B

P(B)= Marginal probability of B.

If the probability of A is given and we need to find the probability of B, then it will be given as:

$$P(B|A) = \frac{P(A \wedge B)}{P(A)}$$

It can be explained by using the below Venn diagram, where B is occurred event, so sample space will be reduced to set B, and now we can only calculate event A when event B is already occurred by dividing the probability of **P(A∧*B*) by P(B)**.



Example:

In a group of 100 computer buyers, 40 bought CPU, 30 purchased monitor, and 20 purchased CPU and monitors. If a computer buyer chose at random and bought a CPU, what is the probability they also bought a Monitor?

Solution: As per the first event, 40 out of 100 bought CPU,

So, P(A) = 40% or 0.4

Now, according to the question, 20 buyers purchased both CPU and monitors. So, this is the intersection of the happening of two events. Hence,

$P(A \cap B) = 20\%$ or 0.2

By the formula of conditional probability we know;

$P(B|A) = P(A \cap B)/P(B)$

$P(B|A) = 0.2/0.4 = 2/4 = \frac{1}{2} = 0.5$

The probability that a buyer bought a monitor, given that they purchased a CPU, is 50%.

Bayes' theorem in Artificial intelligence

Bayes' theorem:

Bayes' theorem is also known as **Bayes' rule, Bayes' law**, or **Bayesian reasoning**, which determines the probability of an event with uncertain knowledge.

In probability theory, it relates the conditional probability and marginal probabilities of two random events.

Bayes' theorem was named after the British mathematician **Thomas Bayes**. The **Bayesian inference** is an application of Bayes' theorem, which is fundamental to Bayesian statistics.

It is a way to calculate the value of $P(B|A)$ with the knowledge of $P(A|B)$.

Bayes' theorem allows updating the probability prediction of an event by observing new information of the real world.

Example: If cancer corresponds to one's age then by using Bayes' theorem, we can determine the probability of cancer more accurately with the help of age.

Bayes' theorem can be derived using product rule and conditional probability of event A with known event B:

As from product rule we can write:

1. $P(A \wedge B)= P(A|B) P(B)$ or

Similarly, the probability of event B with known event A:

1. $P(A \wedge B)= P(B|A) P(A)$

Equating right hand side of both the equations, we will get:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \qquad \text{....(a)}$$

The above equation (a) is called as **Bayes' rule** or **Bayes' theorem**. This equation is basic of most modern AI systems for **probabilistic inference**.

It shows the simple relationship between joint and conditional probabilities. Here,

P(A|B) is known as **posterior**, which we need to calculate, and it will be read as Probability of hypothesis A when we have occurred an evidence B.

P(B|A) is called the likelihood, in which we consider that hypothesis is true, then we calculate the probability of evidence.

P(A) is called the **prior probability**, probability of hypothesis before considering the evidence

P(B) is called **marginal probability**, pure probability of an evidence.

In the equation (a), in general, we can write P (B) = P(A)*P(B|Ai), hence the Bayes' rule can be written as:

$$P(A_i|B) = \frac{P(A_i)*P(B|A_i)}{\sum_{i=1}^{k} P(A_i)*P(B|A_i)}$$

Where A_1, A_2, A_3,........, A_n is a set of mutually exclusive and exhaustive events.

Applying Bayes' rule:

Bayes' rule allows us to compute the single term P(B|A) in terms of P(A|B), P(*B*), and P(A). This is very useful in cases where we have a good probability of these three terms and want to determine the fourth one. Suppose we want to perceive the effect of some unknown cause, and want to compute that cause, then the Bayes' rule becomes:

Example:

Question: From a standard deck of playing cards, a single card is drawn. The probability that the card is king is 4/52, then calculate posterior probability P(King|Face), which means the drawn face card is a king card.

Solution:

$$\mathbf{P(king|face)} = \frac{P(Face|king) * P(King)}{P(Face)} \quad ....(i)$$

P(king): probability that the card is King= 4/52= 1/13

P(face): probability that a card is a face card= 3/13

P(Face|King): probability of face card when we assume it is a king = 1

Putting all values in equation (i) we will get:

$$P(king|face) = \frac{1 * (\frac{1}{13})}{(\frac{3}{13})} = \mathbf{1/3,} \text{ \textbf{it is a probability that a face card is a king card.}}$$

Application of Bayes' theorem in Artificial intelligence:

Following are some applications of Bayes' theorem:

- It is used to calculate the next step of the robot when the already executed step is given.
- Bayes' theorem is helpful in weather forecasting.
- It can solve the Monty Hall problem.

Bayesian Belief Network in artificial intelligence

Bayesian belief network is key computer technology for dealing with probabilistic events and to solve a problem which has uncertainty. We can define a Bayesian network as:

"A Bayesian network is a probabilistic graphical model which represents a set of variables and their conditional dependencies using a directed acyclic graph."

It is also called a **Bayes network, belief network, decision network**, or **Bayesian model**.

Bayesian networks are probabilistic, because these networks are built from a **probability distribution**, and also use probability theory for prediction and anomaly detection.

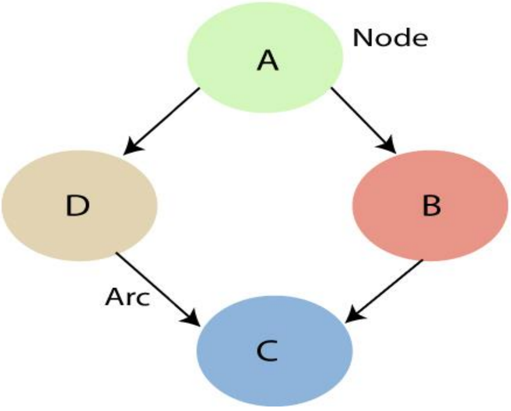
Real world applications are probabilistic in nature, and to represent the relationship between multiple events, we need a Bayesian network. It can also be used in various tasks including **prediction, anomaly detection, diagnostics, automated insight, reasoning, time series prediction**, and **decision making under uncertainty**.

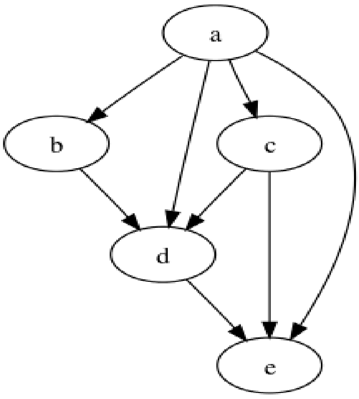
Bayesian Network can be used for building models from data and experts opinions, and it consists of two parts:

- **Directed Acyclic Graph**
- **Table of conditional probabilities.**

The generalized form of Bayesian network that represents and solve decision problems under uncertain knowledge is known as an **Influence diagram**.

A Bayesian network graph is made up of nodes and Arcs (directed links), where:





- Each **node** corresponds to the random variables, and a variable can be **continuous** or **discrete**.
- **Arc or directed arrows** represent the causal relationship or conditional probabilities between random variables. These directed links or arrows connect the pair of nodes in the graph.

These links represent that one node directly influence the other node, and if there is no directed link that means that nodes are independent with each other

- **In the above diagram, A, B, C, and D are random variables represented by the nodes of the network graph.**
- **If we are considering node B, which is connected with node A by a directed arrow, then node A is called the parent of Node B.**
- **Node C is independent of node A.**

The Bayesian network has mainly two components:

- **Causal Component**
- **Actual numbers**

Each node in the Bayesian network has condition probability distribution **$P(X_i |Parent(X_i))$**, which determines the effect of the parent on that node.

Bayesian network is based on Joint probability distribution and conditional probability. So let's first understand the joint probability distribution:

Joint probability distribution:

If we have variables x1, x2, x3,....., xn, then the probabilities of a different combination of x1, x2, x3.. xn, are known as Joint probability distribution.

$P[x_1, x_2, x_3,....., x_n]$, it can be written as the following way in terms of the joint probability distribution.

$= P[x_1| x_2, x_3,....., x_n]P[x_2, x_3,....., x_n]$

$= P[x_1| x_2, x_3,....., x_n]P[x_2|x_3,....., x_n]....P[x_{n-1}|x_n]P[x_n].$

In general for each variable Xi, we can write the equation as:

$$P(X_i|X_{i-1},........., X_1) = P(X_i |Parents(X_i))$$

Explanation of Bayesian network:

Let's understand the Bayesian network through an example by creating a directed acyclic graph:

Example: Harry installed a new burglar alarm at his home to detect burglary.

The alarm reliably responds at detecting a burglary but also responds for minor earthquakes.

Harry has two neighbors David and Sophia, who have taken a responsibility to inform Harry at work when they hear the alarm.

David always calls Harry when he hears the alarm, but sometimes he got confused with the phone ringing and calls at that time too.

On the other hand, Sophia likes to listen to high music, so sometimes she misses to hear the alarm.

Here we would like to compute the probability of Burglary Alarm.

Problem:

Calculate the probability that alarm has sounded, but there is neither a burglary, nor an earthquake occurred, and David and Sophia both called the Harry.

Solution:

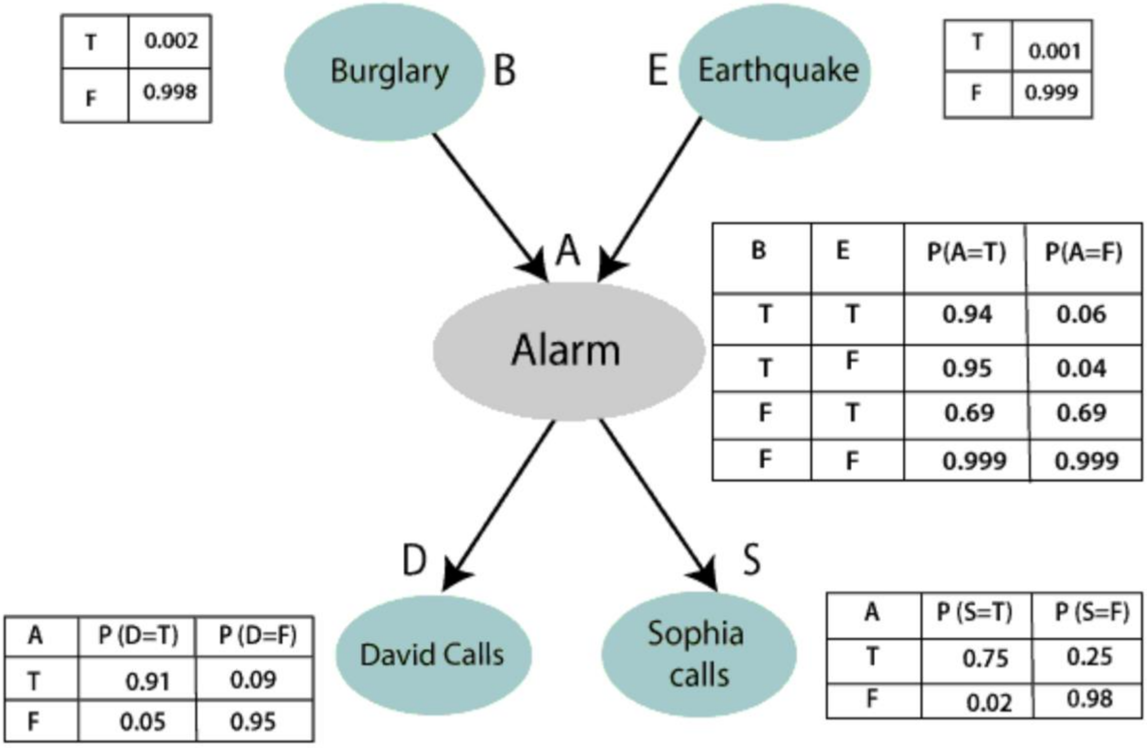
- The Bayesian network for the above problem is given below. The network structure is showing that burglary and earthquake is the parent node of the alarm and directly affecting the probability of alarm's going off, but David and Sophia's calls depend on alarm probability.
- The network is representing that our assumptions do not directly perceive the burglary and also do not notice the minor earthquake, and they also not confer before calling.
- The conditional distributions for each node are given as conditional probabilities table or CPT.
- Each row in the CPT must be sum to 1 because all the entries in the table represent an exhaustive set of cases for the variable.

- In CPT, a boolean variable with k boolean parents contains 2^K probabilities. Hence, if there are two parents, then CPT will contain 4 probability values

List of all events occurring in this network:

- **Burglary (B)**
- **Earthquake(E)**
- **Alarm(A)**
- **David Calls(D)**
- **Sophia calls(S)**

We can write the events of problem statement in the form of probability: **P[D, S, A, B, E]**



T	0.002
F	0.998

T	0.001
F	0.999

B	E	P(A=T)	P(A=F)
T	T	0.94	0.06
T	F	0.95	0.04
F	T	0.69	0.69
F	F	0.999	0.999

A	P (D=T)	P (D=F)
T	0.91	0.09
F	0.05	0.95

A	P (S=T)	P (S=F)
T	0.75	0.25
F	0.02	0.98

Let's take the observed probability for the Burglary and earthquake component:

P(B= True) = 0.002, which is the probability of burglary.

P(B= False)= 0.998, which is the probability of no burglary.

P(E= True)= 0.001, which is the probability of a minor earthquake

P(E= False)= 0.999, Which is the probability that an earthquake not occurred.

We can provide the conditional probabilities as per the below tables:

Conditional probability table for Alarm A:

The Conditional probability of Alarm A depends on Burglar and earthquake:

B	E	P(A= True)	P(A= False)
True	True	0.94	0.06
True	False	0.95	0.04
False	True	0.31	0.69
False	False	0.001	0.999

Conditional probability table for David Calls:

The Conditional probability of David that he will call depends on the probability of Alarm.

A	P(D= True)	P(D= False)
True	0.91	0.09
False	0.05	0.95

Conditional probability table for Sophia Calls:

The Conditional probability of Sophia that she calls is depending on its Parent Node "Alarm."

A	P(S= True)	P(S= False)

True	0.75	0.25
False	0.02	0.98

From the formula of joint distribution, we can write the problem statement in the form of probability distribution:

P(S, D, A, ¬B, ¬E) = P (S|A) *P (D|A)*P (A|¬B ^ ¬E) *P (¬B) *P (¬E).

= 0.75* 0.91* 0.001* 0.998*0.999

= 0.00068045.

Hence, a Bayesian network can answer any query about the domain by using Joint distribution.

Exact inference

Find posterior/conditional probability for any given query.

- X - query can be on single/multi variables
- E - set of evidences (E_1, E_2,....E_m)
- Y - hidden variable that are neither evidence nor query (Y_1,Y_2,.....,Y_L).

  Thus, the complete set of variables is X = {X} ∪ E ∪ Y
  A typical query asks for the posterior probability distribution P(X | e).

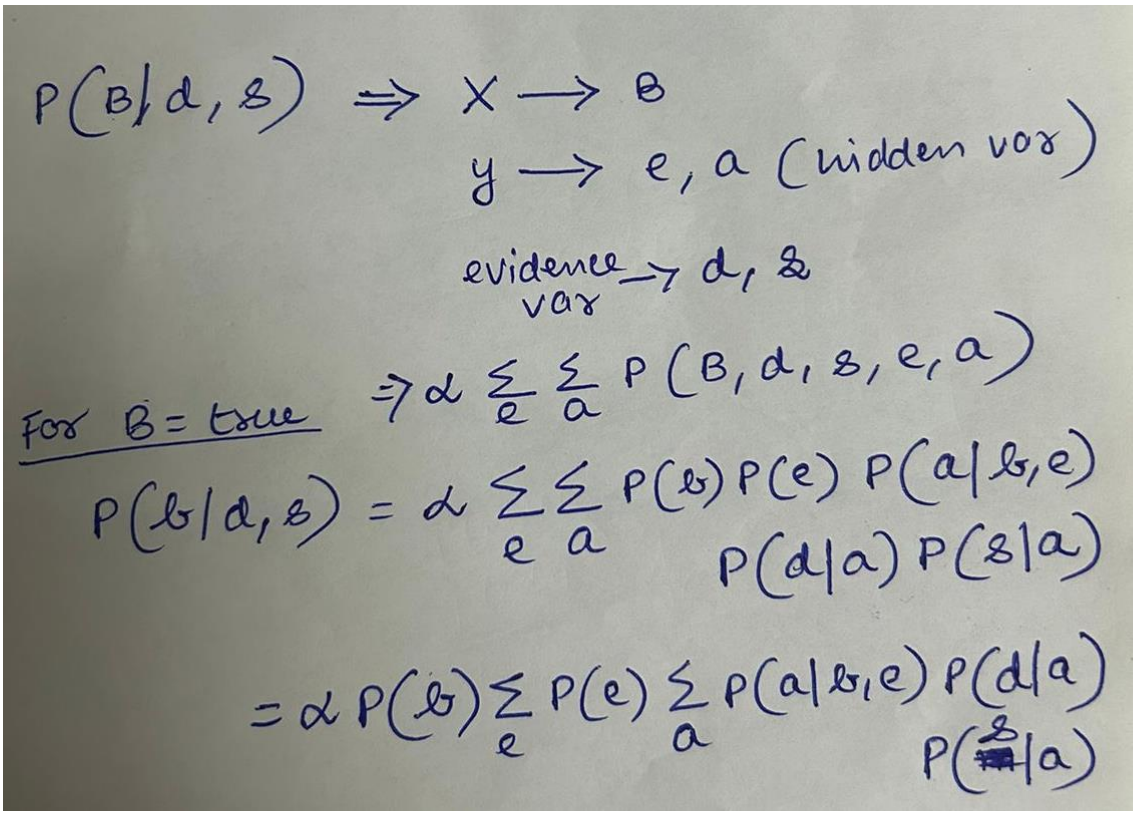
  Eg. P(Burglary/John Calls= True, Mary Calls= True)

1) Inference by enumeration

- any conditional probability can be computed by summing terms from the full joint distribution.
- a query **P(X | e)** can be answered using

  P (X|e) = $\alpha P(X,e) = \alpha \sum_y P(X,e,y)$

- query can be answered using a Bayesian network by computing sums of products of conditional probabilities from the network
- $\sum \rightarrow +$



$P(B/d,s) \Rightarrow X \rightarrow B$

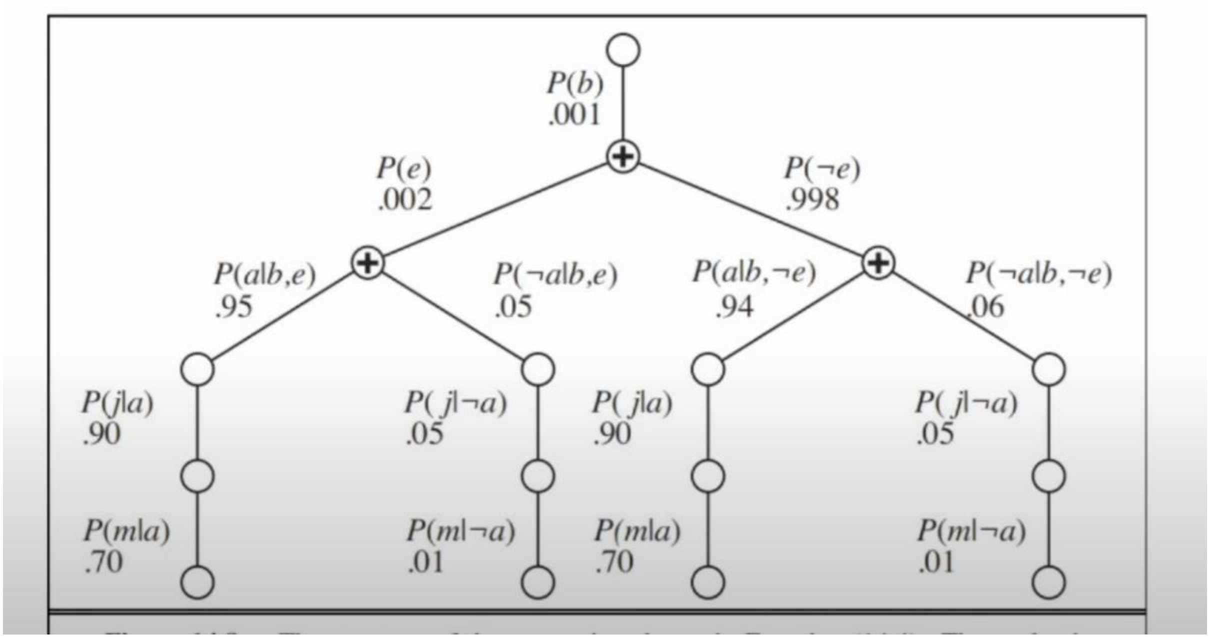
$y \rightarrow e, a$ (hidden var)

evidence var → d, s

For B = true $\Rightarrow \alpha \sum_{e} \sum_{a} P(B,d,s,e,a)$

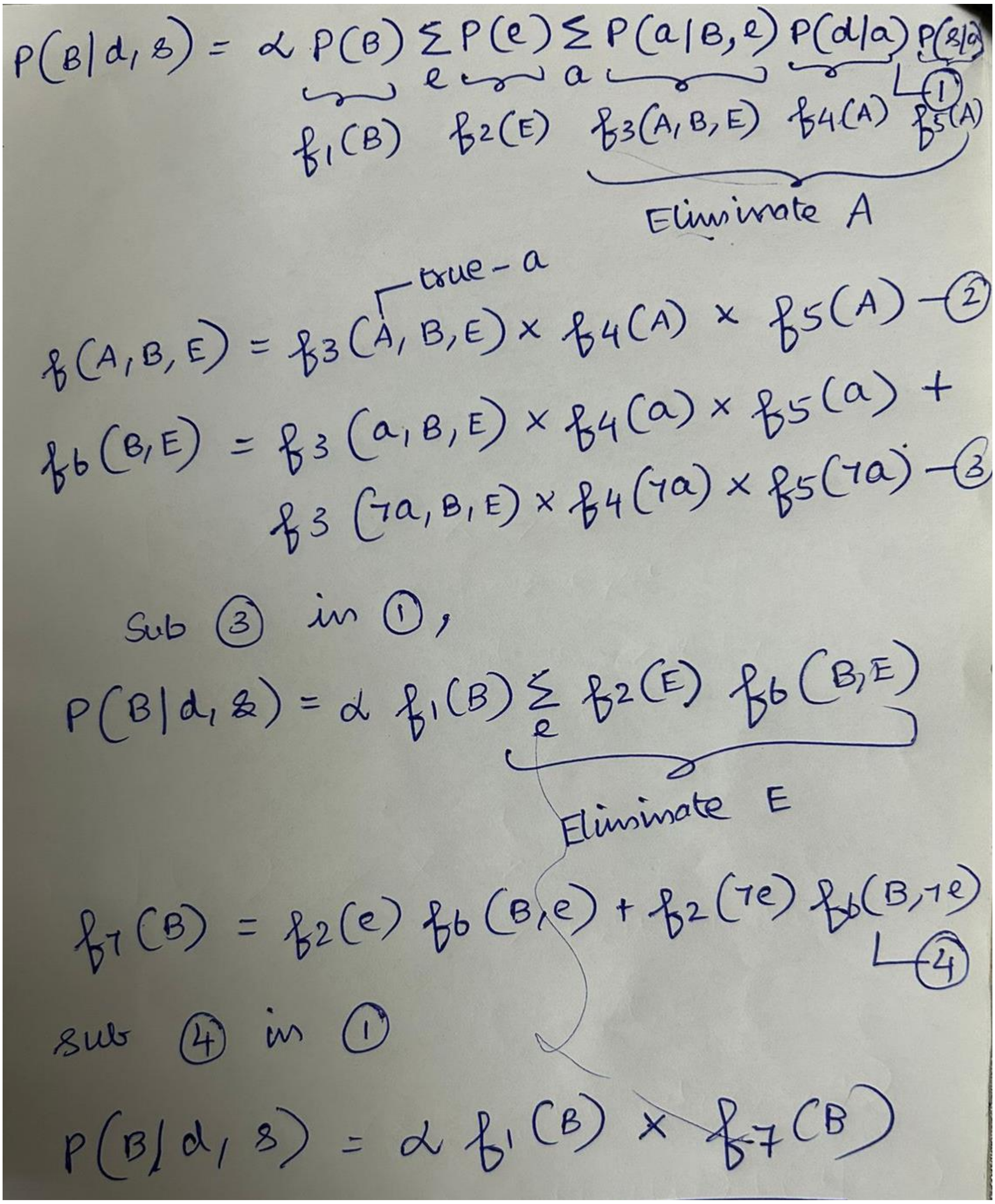
$$P(b/d,s) = \alpha \sum_{e} \sum_{a} P(b)P(e)P(a|b,e) P(d|a)P(s|a)$$

$$= \alpha P(b) \sum_{e} P(e) \sum_{a} P(a|b,e) P(d|a) P(s|a)$$



2) Variable Elimination Algorithm

- **To avoid redundancy**
- **Right to left computation**
- **Intermediatory repetitive results are stored**



$$P(B|d,s) = \alpha \underbrace{P(B)}_{f_1(B)} \sum_e \underbrace{P(e)}_{f_2(E)} \underbrace{\sum_a \underbrace{P(a|B,e)}_{f_3(A,B,E)} \underbrace{P(d|a)}_{f_4(A)} \underbrace{P(s|a)}_{f_5(A)}}_{\text{Eliminate A}} \quad ①$$

$$f(A,B,E) = f_3(\overset{\text{true - a}}{A}, B, E) \times f_4(A) \times f_5(A) \quad ②$$

$$f_6(B,E) = f_3(a,B,E) \times f_4(a) \times f_5(a) + f_3(\neg a,B,E) \times f_4(\neg a) \times f_5(\neg a) \quad ③$$

Sub ③ in ①,

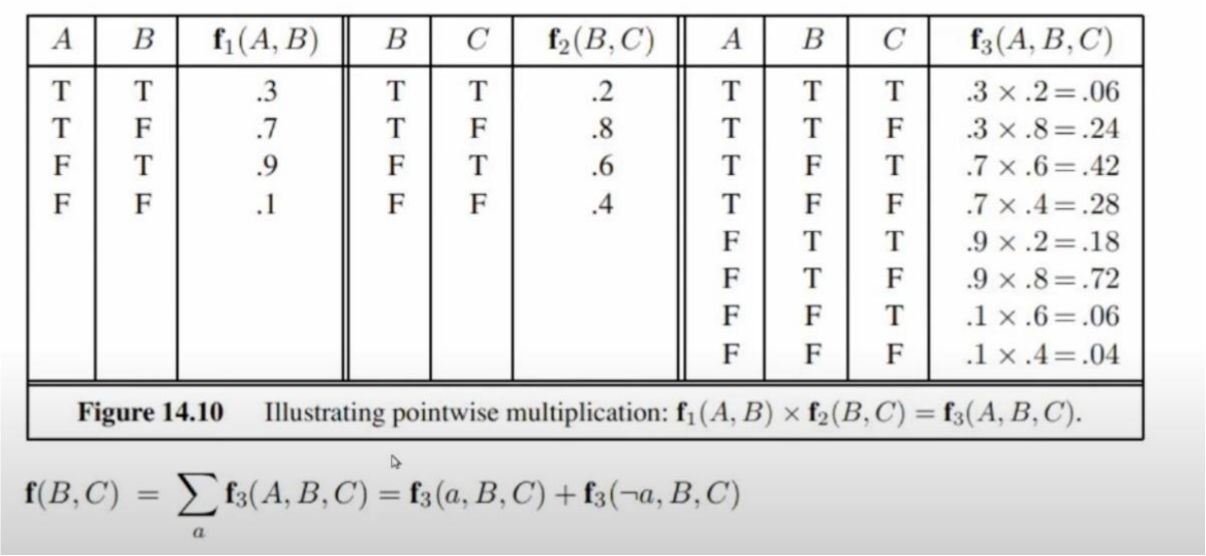
$$P(B|d,s) = \alpha f_1(B) \sum_e \underbrace{f_2(E)\ f_6(B,E)}_{\text{Eliminate E}}$$

$$f_7(B) = f_2(e)\ f_6(B,e) + f_2(\neg e)\ f_6(B,\neg e) \quad ④$$

sub ④ in ①

$$P(B|d,s) = \alpha f_1(B) \times f_7(B)$$

•



A	B	$\mathbf{f}_1(A,B)$	B	C	$\mathbf{f}_2(B,C)$	A	B	C	$\mathbf{f}_3(A,B,C)$
T	T	.3	T	T	.2	T	T	T	$.3 \times .2 = .06$
T	F	.7	T	F	.8	T	T	F	$.3 \times .8 = .24$
F	T	.9	F	T	.6	T	F	T	$.7 \times .6 = .42$
F	F	.1	F	F	.4	T	F	F	$.7 \times .4 = .28$
						F	T	T	$.9 \times .2 = .18$
						F	T	F	$.9 \times .8 = .72$
						F	F	T	$.1 \times .6 = .06$
						F	F	F	$.1 \times .4 = .04$

Figure 14.10 Illustrating pointwise multiplication: $\mathbf{f}_1(A,B) \times \mathbf{f}_2(B,C) = \mathbf{f}_3(A,B,C)$.

$$\mathbf{f}(B,C) = \sum_a \mathbf{f}_3(A,B,C) = \mathbf{f}_3(a,B,C) + \mathbf{f}_3(\neg a,B,C)$$

Approximate Inference

What is approximate inference?

- It is a method of estimating probabilities in Bayesian networks also called ‘Monte Carlo’ algorithms.

- We will discuss two types of algorithms: ***Direct sampling*** and ***Markov chain sampling.***

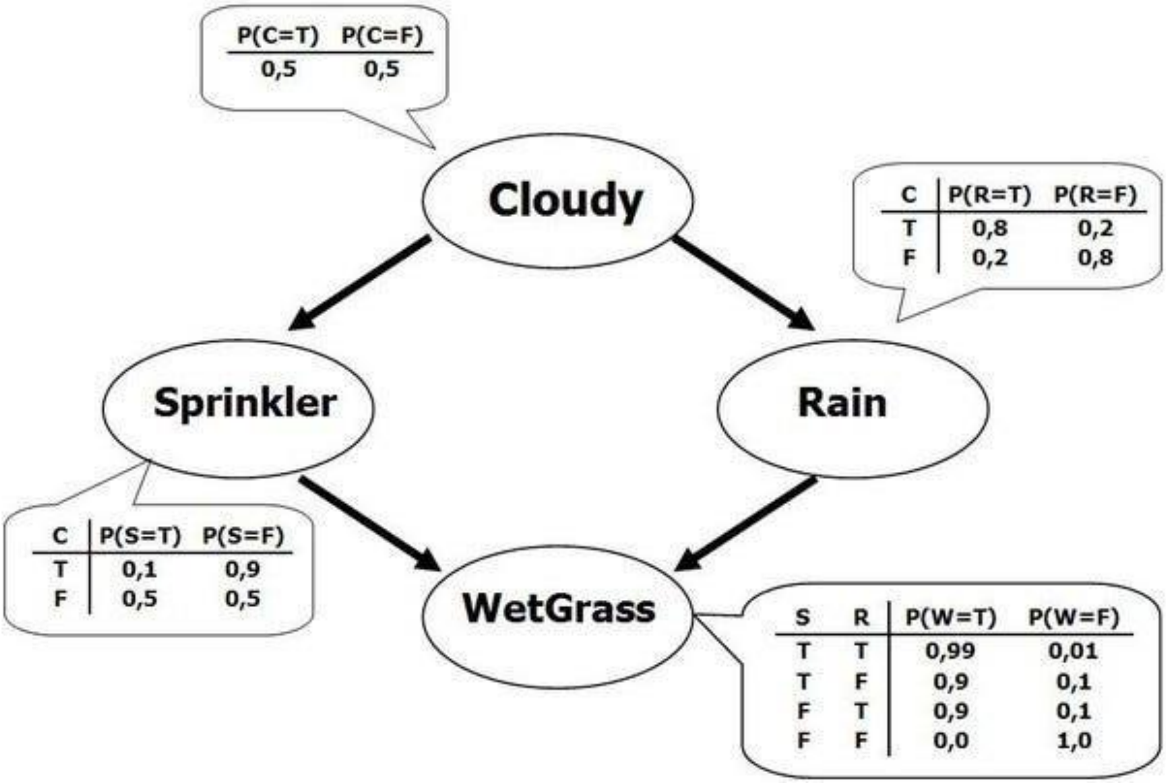
Why use approximate inference?

- Exact inference becomes intractable for large multiply-connected networks
- Variable elimination can have exponential time and space complexity
- Exact inference is strictly HARDER than NP-complete problems (#P-hard)

Direct Sampling

In direct sampling we take samples of events. We expect the frequency of the samples to converge on the probability of the event.

Example:



P(C=T)	P(C=F)
0,5	0,5

C	P(R=T)	P(R=F)
T	0,8	0,2
F	0,2	0,8

C	P(S=T)	P(S=F)
T	0,1	0,9
F	0,5	0,5

S	R	P(W=T)	P(W=F)
T	T	0,99	0,01
T	F	0,9	0,1
F	T	0,9	0,1
F	F	0,0	1,0

Prior sampling

Consider Variables -> C, S, R, W

Sample each variable in some order

P (cloudy C) = T- 0.5, F- 0.5 => T-0.5

P (S/C = True) = T – 0.1, F- 0.9 => F-0.9

P (R/C = True) = T – 0.8, F – 0.2 => T-0.8

P (W/S = False, R=True) = T-0.9, F-0.1 => T – 0.9

Probability of a specific event = 0.5 x 0.9 x 0.8 x 0.9 = 32.4 %

Rejection Sampling

- Used to compute conditional probabilities P(X|e)

- Generate samples as before
- Reject samples that do not match evidence
- Estimate by counting the how often event X is in the resulting samples

For example,

Find P(R/S = True) using 100 samples, N = 100

S = False -> 73 samples. (Rejected)

S = True -> 27 samples => R = True (8 samples)

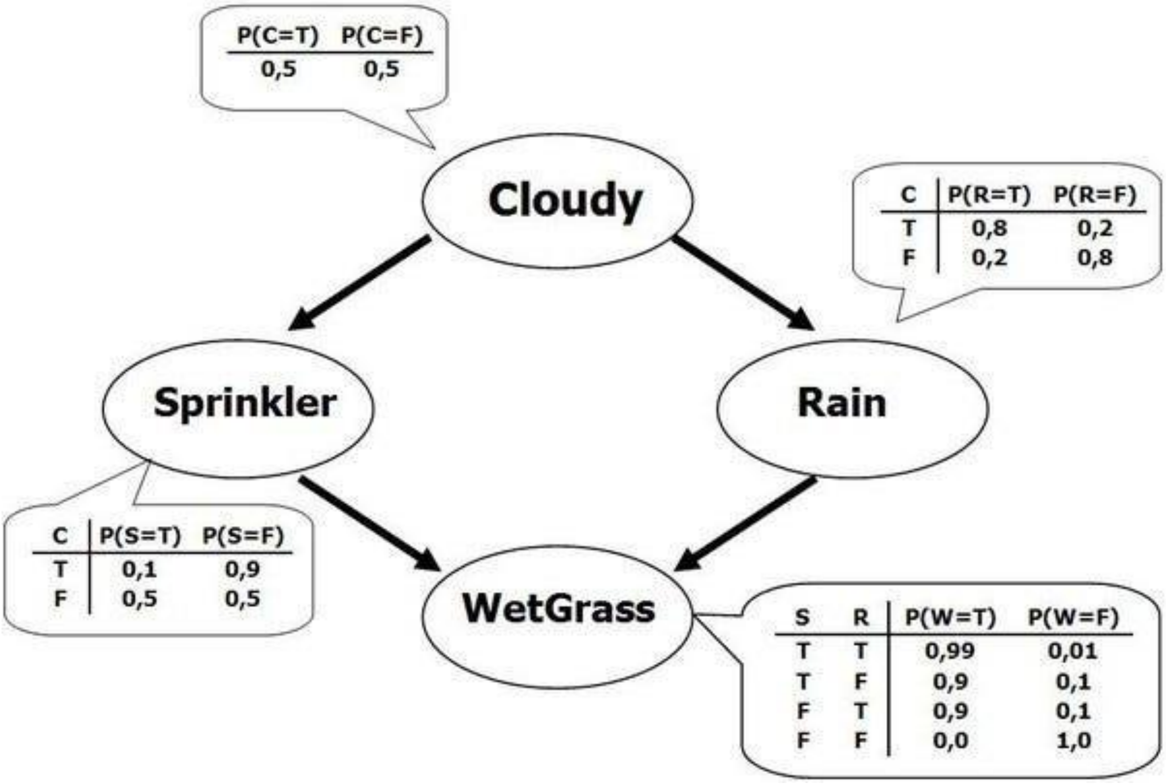
=> R = False (19 samples) (Rejected)

P (R/S = True) = Normalize (T-8, F-19) = 8/27, 19/27 = 0.296, 0.704

Drawback – rejects so many samples if the evidence condition matches

Likelihood Weighting

- Avoid inefficiency of rejection sampling
- Fix values for evidence variables and only sample the remaining variables
- Weight samples with regard to how likely they are



P(C=T)	P(C=F)
0,5	0,5

C	P(R=T)	P(R=F)
T	0,8	0,2
F	0,2	0,8

C	P(S=T)	P(S=F)
T	0,1	0,9
F	0,5	0,5

S	R	P(W=T)	P(W=F)
T	T	0,99	0,01
T	F	0,9	0,1
F	T	0,9	0,1
F	F	0,0	1,0

Example:

Consider the query P (R/C = true; W = true)

(C, W - evidence variables, R, S – non evidence variable)

Weight must be considered only for the evidence variables

Assume Weight Wt = 1

1) P (C=true) = 0.5
   Wt = Wt x P (C=true) = 1 x 0.5 = 0.5
2) P (S/C = true) = T-0.1, **F-0.9** => F-0.9
3) P (R/C = true) = **T-0.8**, F-0.2 = >T-0.8
4) P (W = true) = 0.9
   Wt = Wt x P (W=true) = 0.5 x 0.9 = 0.45

Markov Chain Sampling

- Generate events by making a random change to the preceding event
- This change is made using the Markov Blanket of the variable to be changed
- Markov Blanket = parents, children, children's parents
- Tally and normalize results

Consider the example below,

Cloudy	Rain	Sprinkler	Wetgrass
T	T	F	T
F	T	F	T
F	T	F	***F***

Gibbs Sampling

- The Gibbs sampling algorithm for Bayesian networks starts with an arbitrary state (with the evidence variables fixed at their observed values)
- Generates a next state by **randomly sampling a value for one of the non-evidence variables** $X_{i.}$
- The sampling for X_i is done conditioned on the current values of the variables in the Markov blanket of X_i

Example

1) Initial state – random

   P (R/ S = True ; W = True) evidence-> S, W & non evidence -> R, C

Cloudy	Sprinkler	Rain	Wetgrass
T	T	F	T

2) Sample non-evidence (Cloudy and rain)

Cloudy	Sprinkler	Rain	Wetgrass
F	T	F	T
F	T	***T***	T

N = 80, R = True => 20 & R== False => 60

- Markov blanket

  P(C/S = True ; R = False) => False