

→ Chain Rule → Joint = Marginal $\times$ Conditional

→ How to get conditional given Joint

$$P(Y/x) = \frac{P(Y, x)}{P(x)} = \frac{\text{Conditional}}{\text{Marginal}}$$

↓ Normalising

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

T	W	P
cold	sun	0.2
cold	rain	0.3

$$P(W|T=c)$$

W	P
sun	0.4
rain	0.6

$$P(x_1|x_2) = \frac{P(x_1, x_2)}{P(x_2)} = \frac{P(x_1, x_2)}{\sum_{x_1} P(x_1, x_2)}$$
$$\frac{P(W=\text{sun}, T=c)}{P(T=c)}$$
$$p(X = x_i) = \frac{c_i}{N}.$$
$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$
$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

C_i

X value is given

- Conditional probability query: $P(x|e)$
- Maximum a posteriori probability: What value of x maximizes $P(x|e)$?
- General question: What's the whole probability distribution over variable X given evidence e — what is $P(X|e)$?

- Derived probability from known probability
- Usually from joint distribution
- Enumeration → exact, exponential complexity
- Variable elimination → exact, worst case exponential complexity
- Sampling (approximate)

- General case:
 - Evidence variables: $E_1 \dots E_k = e_1 \dots e_k$
 - Query* variable: Q
 - Hidden variables: $H_1 \dots H_r$

$\left. \begin{array}{l} X_1, X_2, \dots, X_n \\ \text{All variables} \end{array} \right\}$

* Works fine with multiple query variables, too

$$P(Q|e_1 \dots e_k)$$

$$p(Q|e_1, e_2, \dots, e_k) = \frac{p(Q, e_1, e_2, \dots, e_k)}{Z}$$

* Works fine with multiple query variables, too

$$P(Q|e_1 \dots e_k)$$

- Step 2: Sum out H to get joint of Query and evidence

$$\times \frac{1}{Z}$$

$$Z = \sum_{e_1 \cdots e_k} P(Q, e_1 \cdots e_k) = \frac{1}{Z} P(Q, e_1 \cdots e_k)$$

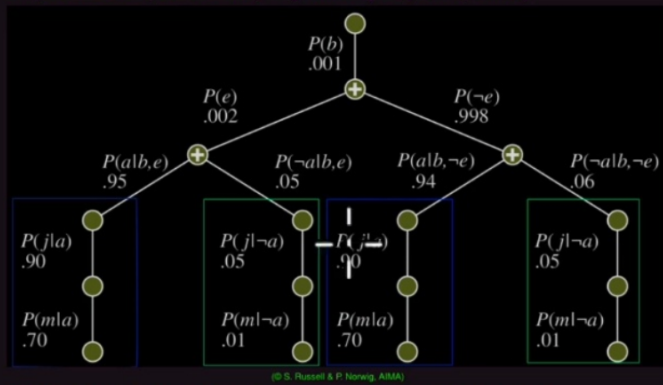
- Given a bayesian N/w → Factorise
- Use Summation over hidden variables (unobserved)
- Calculate Normalising value
- If there are n of unobserved (hidden) variable with each can take d values
⇒ d^n terms in summation
- d^n → exponential → Too much costly

Normalisation is not calculated explicitly,
Once the numerator is found out, we just normalise

$$P(b|j,m) = \alpha P(b) \sum_e P(e) \sum_a P(a|b,e) P(j|a) P(m|a) = \alpha \cdot 0.00059224$$

$$P(\neg b|j,m) = \alpha P(\neg b) \sum_e P(e) \sum_a P(a|\neg b,e) P(j|a) P(m|a) = \alpha \cdot 0.0014919$$

$$\Rightarrow P(b|j,m) = \alpha \cdot (0.00059224, 0.0014919) = [\text{normal}] \approx (0.284, 0.716)$$



Repeated computation: $P(j|a)P(m|a)$ & $P(j|\neg a)P(m|\neg a)$ for each value of e

$$P(+b)P(+e)P(+a|+b,+e)P(+j|+a)P(+m|+a) +$$

$$P(+b)P(+e)P(-a|+b,+e)P(+j|-a)P(+m|-a) +$$

$$P(+b)P(-e)P(+a|+b,-e)P(+j|+a)P(+m|+a) +$$

$$P(+b)P(-e)P(-a|+b,-e)P(+j|-a)P(+m|-a)$$

Because E and A are unobserved, we have to consider all cases

2 unobserved each takes 2 values $\Rightarrow 2^2 = 4$ terms in summation

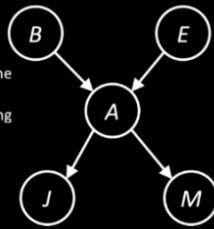
$\rightarrow$ Disadvantage of enumeration is this repeated redundant calculation

Inference by Enumeration in Bayes' Net

Given unlimited time, inference in BNs is easy

Reminder of inference by enumeration:

- Any probability of interest can be computed by summing entries from the joint distribution: $P(Q|e) = \alpha \sum_h P(Q,h,e)$
- Entries from the joint distribution can be obtained from a BN by multiplying the corresponding conditional probabilities



$P(W=\text{sun})? \Rightarrow P(S,T,P,W=\text{sun})$

$P(W=\text{sun} | S=\text{winter})? \Rightarrow \frac{P(W=\text{sun}, S=\text{winter})}{P(S=\text{winter})}$

$P(W=\text{sun} | S=\text{winter}, T=\text{hot})?$

$\frac{P(W=\text{sun}, S=\text{winter}, T=\text{hot})}{P(S=\text{winter}, T=\text{hot})} \rightarrow \text{Normalise}$

S	T	W	P
summer	hot	sun	0.30
summer	hot	rain	0.05
summer	cold	sun	0.10
summer	cold	rain	0.05
winter	hot	sun	0.10
winter	hot	rain	0.05
winter	cold	sun	0.15
winter	cold	rain	0.20

Inference by Enumeration

Normalization Constant (computed after the fact)

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

Query Variables

Evidence Variables

Hidden Variables

$P(W)?$

$$p(W=\text{sun}) = 0.3 + 0.1 + 0.1 + 0.15 = 0.65$$

$$p(W=\text{rain}) = 0.05 + 0.05 + 0.05 + 0.2 = 0.35$$

$P(W | \text{winter})?$

$$p(W=\text{sun}, \text{winter}) = 0.1 + 0.15 = 0.25$$

$$p(W=\text{rain}, \text{winter}) = 0.05 + 0.2 = 0.25$$

$$p(W=\text{sun} | \text{winter}) = 0.25 / (0.25 + 0.25) = 0.5$$

$$p(W=\text{rain} | \text{winter}) = 0.25 / (0.25 + 0.25) = 0.5$$

$P(W | \text{winter}, \text{hot})?$

$$p(W=\text{sun}, \text{winter}, \text{hot}) = 0.1$$

$$p(W=\text{rain}, \text{winter}, \text{hot}) = 0.05$$

$$p(W=\text{sun} | \text{winter}, \text{hot}) = 0.1 / (0.1 + 0.05) = 2/3$$

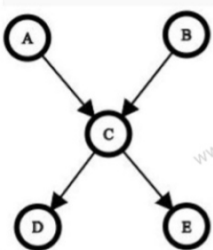
$$p(W=\text{rain} | \text{winter}, \text{hot}) = 0.05 / (0.1 + 0.05) = 1/3$$

S	T	W	P
summer	hot	sun	0.30
summer	hot	rain	0.05
summer	cold	sun	0.10
summer	cold	rain	0.05
winter	hot	sun	0.10
winter	hot	rain	0.05
winter	cold	sun	0.15
winter	cold	rain	0.20

$P(W, \text{winter}) = 0.25$

$P(W | \text{winter}) = \frac{0.25}{0.25 + 0.25} = 0.5$

9) Consider the following Bayesian Network with the given information. What is the value of $P(\neg a | d, e)$ up to 3 decimal places?



$P(a) = 0.001, P(b) = 0.002$
 $P(c|a,b) = 0.95, P(c|a,\neg b) = 0.94, P(c|\neg a,b) = 0.29, P(c|\neg a,\neg b) = 0.001$
 $P(d|c) = 0.9, P(d|\neg c) = 0.05$
 $P(e|c) = 0.7, P(e|\neg c) = 0.01$

$$P(A,B,C,D,E) = P(A) \cdot P(B) \cdot P(C|A,B) \cdot P(D|C) \cdot P(E|C)$$

$$P(\neg A | D, E) = \frac{P(\neg A, D, E)}{P(D, E)}$$

$$\sum_{B,C} P(\neg A) \cdot P(B) \cdot P(C|A,B) \cdot P(D|C) \cdot P(E|C)$$

$$= 0.999 \times 0.002 \times 0.29 \times 0.9 \times 0.7 +$$

$$0.999 \times 0.998 \times 0.001 \times 0.9 \times 0.7 +$$

$$0.999 \times 0.002 \times 0.71 \times 0.05 \times 0.01 +$$

$$0.999 \times 0.998 \times 0.999 \times 0.05 \times 0.01$$

$B=1, C=1$
 $B=0, C=1$
 $B=1, C=0$
 $B=0, C=0$

Variable Elimination

Can we do better?

- Consider $uwv + uwz + uxy + uxz + vwy + vwz + vxy + vxz$
 - 16 multiplies, 7 adds
 - Lots of repeated subexpressions!
- Rewrite as $(u+v)(w+x)(y+z)$
 - 2 multiplies, 3 adds

$$\sum_{e,a} P(B) P(e) P(a|B,e) P(j|a) P(m|a)$$

$$= P(B)P(e)P(a|B,e)P(j|a)P(m|a) + P(B)P(-e)P(a|B,-e)P(j|a)P(m|a)$$

$$+ P(B)P(e)P(-a|B,e)P(j|-a)P(m|-a) + P(B)P(-e)P(-a|B,-e)P(j|-a)P(m|-a)$$

Lots of repeated subexpressions!

In general, when we write $P(Y_1 \dots Y_N | X_1 \dots X_M)$

- It is a "factor," a multi-dimensional array (matrix)
- Its values are $P(y_1 \dots y_N | x_1 \dots x_M)$

Inference by Enumeration: Procedural Outline

- Track objects called **factors**
- Initial factors are local CPTs (one per node)

$P(R)$	$P(T R)$	$P(L T)$
$\begin{matrix} +r & 0.1 \\ -r & 0.9 \end{matrix}$	$\begin{matrix} +r & +t & 0.8 \\ +r & -t & 0.2 \\ -r & +t & 0.1 \\ -r & -t & 0.9 \end{matrix}$	$\begin{matrix} +t & +l & 0.3 \\ +t & -l & 0.7 \\ -t & +l & 0.1 \\ -t & -l & 0.9 \end{matrix}$

- Any known values are selected

E.g. if we know $L = +l$ the initial factors are

$P(R)$	$P(T R)$	$P(+l T)$
$\begin{matrix} +r & 0.1 \\ -r & 0.9 \end{matrix}$	$\begin{matrix} +r & +t & 0.8 \\ +r & -t & 0.2 \\ -r & +t & 0.1 \\ -r & -t & 0.9 \end{matrix}$	$\begin{matrix} +t & +l & 0.3 \\ -t & +l & 0.1 \end{matrix}$

Procedure: Join all factors, eliminate all hidden variables, normalize

Variable Elimination: The basic ideas

- Move summations inwards as far as possible

$$P(B | j, m) = \alpha \sum_{e,a} P(B) P(e) P(a|B,e) P(j|a) P(m|a)$$

$$= \alpha P(B) \sum_e P(e) \sum_a P(a|B,e) P(j|a) P(m|a)$$

- Do the calculation from the inside out
 - i.e., sum over a first, then sum over e

Joint distribution: $P(X,Y)$

- Entries $P(x,y)$ for all x, y
- $|X| \times |Y|$ matrix
- Sums to 1

$P(A,J)$

$A \setminus J$	true	false
true	0.09	0.01
false	0.045	0.855

$$|A| = 2$$

$$|J| = 2$$

Projected joint: $P(x,Y)$

- A slice of the joint distribution
- Entries $P(x,y)$ for one x , all y
- $|Y|$ -element vector
- Sums to $P(x)$

$P(a,J) = P_a(J)$

$A \setminus J$	true	false
true	0.09	0.01

Single conditional: $P(Y | x)$

- Entries $P(y | x)$ for fixed x , all y
- Sums to 1

$P(J|a)$

$A \setminus J$	true	false
true	0.9	0.1

Family of conditionals: $P(X | Y)$

- Multiple conditionals
- Entries $P(x | y)$ for all x, y
- Sums to $|Y|$

$P(J|A)$

$A \setminus J$	true	false
true	0.9	0.1
false	0.05	0.95

Operation 1: Join Factors

First basic operation: **joining factors**

Combining factors:

- Just like a database join
- Get all factors over the joining variable
- Build a new factor over the union of the variables involved

Example: Join on R

$P(R)$	$P(T R)$	$P(R,T)$
$\begin{matrix} +r & 0.1 \\ -r & 0.9 \end{matrix}$	$\begin{matrix} +r & +t & 0.8 \\ +r & -t & 0.2 \\ -r & +t & 0.1 \\ -r & -t & 0.9 \end{matrix}$	$\begin{matrix} +r & +t & 0.08 \\ +r & -t & 0.02 \\ -r & +t & 0.09 \\ -r & -t & 0.81 \end{matrix}$

Computation for each entry: pointwise products $\forall r, t: P(r, t) = P(r) \cdot P(t|r)$

$$P(R,T) \cdot P(L|T) = P(R,T,L)$$

Example: Multiple Joins

$P(R)$	$P(T R)$	$P(L T)$	$P(R,T)$	$P(R,T,L)$
$\begin{matrix} +r & 0.1 \\ -r & 0.9 \end{matrix}$	$\begin{matrix} +r & +t & 0.8 \\ +r & -t & 0.2 \\ -r & +t & 0.1 \\ -r & -t & 0.9 \end{matrix}$	$\begin{matrix} +t & +l & 0.3 \\ +t & -l & 0.7 \\ -t & +l & 0.1 \\ -t & -l & 0.9 \end{matrix}$	$\begin{matrix} +r & +t & 0.08 \\ +r & -t & 0.02 \\ -r & +t & 0.09 \\ -r & -t & 0.81 \end{matrix}$	$\begin{matrix} +r & +t & +l & 0.024 \\ +r & +t & -l & 0.056 \\ +r & -t & +l & 0.002 \\ +r & -t & -l & 0.018 \\ -r & +t & +l & 0.027 \\ -r & +t & -l & 0.063 \\ -r & -t & +l & 0.081 \\ -r & -t & -l & 0.729 \end{matrix}$

Operation 2: Eliminate

Operation 2: Eliminate

Second basic operation: **marginalization**

Take a factor and sum out a variable

- Shrinks a factor to a smaller one
- A projection operation

Example:

$P(R,T)$	sum R	$P(T)$
$\begin{matrix} +r & +t & 0.08 \\ +r & -t & 0.02 \\ -r & +t & 0.09 \\ -r & -t & 0.81 \end{matrix}$		$\begin{matrix} +t & 0.17 \\ -t & 0.83 \end{matrix}$

$$P(+t) = \sum_r P(R, +t)$$

$$= P(+r, +t) + P(-r, +t)$$

$$= 0.08 + 0.09 = 0.17$$

Multiple Elimination

$P(R,T,L)$	Sum out R	$P(T,L)$	Sum out T	$P(L)$
$\begin{matrix} +r & +t & +l & 0.024 \\ +r & +t & -l & 0.056 \\ +r & -t & +l & 0.002 \\ +r & -t & -l & 0.018 \\ -r & +t & +l & 0.027 \\ -r & +t & -l & 0.063 \\ -r & -t & +l & 0.081 \\ -r & -t & -l & 0.729 \end{matrix}$		$\begin{matrix} +t & +l & 0.051 \\ +t & -l & 0.119 \\ -t & +l & 0.083 \\ -t & -l & 0.747 \end{matrix}$		$\begin{matrix} +l & 0.134 \\ -l & 0.886 \end{matrix}$

Thus Far: Multiple Join, Multiple Eliminate (= Inference by Enumeration)

Variable Elimination Algorithm

- Given: BN, evidence e , a query $P(Q|x_e)$
- Choose an **ordering** on variables, e.g., $X_1, \dots, X_n$
- For $i = 1$ to n , If $X_i \notin \{Q, X_e\}$
 - Collect factors $f_1, \dots, f_k$ that include X_i
 - Generate a new factor by eliminating X_i from these factors:

$$g = \sum_{x_i} \prod_{j=1}^k f_j$$

- Multiply all remaining factors
- Normalize $P(Q, x_e)$ to obtain $P(Q|x_e)$

After this summation, X_i is eliminated

Query: $P(X_3|Y_1 = y_1, Y_2 = y_2, Y_3 = y_3)$

Start by inserting evidence, which gives the following initial factors:

$$p(Z)p(X_1|Z)p(X_2|Z)p(X_3|Z)p(y_1|X_1)p(y_2|X_2)p(y_3|X_3)$$

Eliminate X_1 , this introduces the factor $f_1(Z, y_1) = \sum_{x_1} p(x_1|Z)p(y_1|x_1)$, and we are left with:

$$p(Z)f_1(Z, y_1)p(X_2|Z)p(X_3|Z)p(y_2|X_2)p(y_3|X_3)$$

Eliminate X_2 , this introduces the factor $f_2(Z, y_2) = \sum_{x_2} p(x_2|Z)p(y_2|x_2)$, and we are left with:

$$p(Z)f_1(Z, y_1)f_2(Z, y_2)p(X_3|Z)p(y_3|X_3)$$

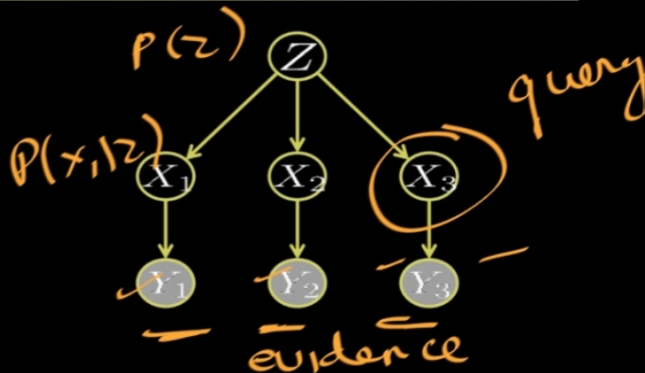
Eliminate Z , this introduces the factor $f_3(y_1, y_2, X_3) = \sum_z p(z)f_1(z, y_1)f_2(z, y_2)p(X_3|z)$, and we are left:

$$p(y_3|X_3), f_3(y_1, y_2, X_3)$$

No hidden variables left. Join the remaining factors to get:

$$f_4(y_1, y_2, y_3, X_3) = P(y_3|X_3)f_3(y_1, y_2, X_3).$$

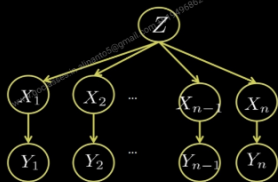
Normalizing over X_3 gives $P(X_3|y_1, y_2, y_3)$.



Computational complexity critically depends on the largest factor being generated in this process. Size of factor = number of entries in table. In example above (assuming binary) all factors generated are of size 2 --- as they all only have one variable (Z, Z , and X_3 respectively).

Variable Elimination Ordering

For the query $P(X_n|y_1, \dots, y_n)$ work through the following two different orderings as done in previous slide: $Z, X_1, \dots, X_{n-1}$ and $X_1, \dots, X_{n-1}, Z$. What is the size of the maximum factor generated for each of the orderings?



In general: the ordering can greatly affect efficiency.

$$P(Z, X_1, X_2, \dots, X_n, Y_1, Y_2, \dots, Y_n) = P(Z) \cdot P(X_1|Z) \cdot P(X_2|Z) \cdot \dots \cdot P(X_n|Z) \cdot P(Y_1|X_1) \cdot P(Y_2|X_2) \cdot \dots \cdot P(Y_n|X_n)$$

a) eliminate $Z, X_1, X_2, \dots, X_{n-1}$ in this order $\rightarrow P(X_n|y_1, y_2, \dots, y_n)$

$$\Rightarrow \sum_Z P(Z) \cdot P(X_1|Z) \cdot P(X_2|Z) \cdot \dots \cdot P(X_n|Z) = f(y_1, y_2, \dots, y_n)$$

$O(d^n)$ $\Leftarrow$ size of factor = d^n assuming binary variables.
 $\rightarrow$ else d^n for d possible values each X_i can take
 $\rightarrow$ This is the largest factor in enumerating

b) eliminate $X_1, X_2, \dots, X_{n-1}, Z$ in this order $\rightarrow P(X_n|y_1, y_2, \dots, y_n)$

$$\Rightarrow \sum_{X_1} P(X_1|Z) \cdot P(Y_1|X_1) = f(Z, y_1)$$

$\rightarrow$ only d is the size
 $\rightarrow$ All factors are like this
 $\rightarrow y_1$ instantiated, $\Rightarrow$ not a variable
 $\rightarrow y_1, y_2 \Rightarrow$ evidence

at the last

$$\Rightarrow \sum_Z P(Z) \cdot f(Z, y_1) \cdot f(Z, y_2) \cdot \dots \cdot f(Z, y_{n-1}) = f(X_n, y_1, y_2, \dots, y_{n-1})$$

$\rightarrow d$ complexity.
 $\rightarrow$ All $y_i \rightarrow y_n$ are evidence & hence not variables.

The computational and space complexity of variable elimination is determined by the largest factor

The elimination ordering can greatly affect the size of the largest factor.

- E.g., previous slide's example 2^n vs. 2

Does there always exist an ordering that only results in small factors?

- No!

You're right to scrutinize that step – good eye. The key point is **evidence**: the Y_i are observed (we're conditioning on $y_1, \dots, y_n$), so they are *instantiated* before/when we run elimination. That changes which variables appear in intermediate factors.

Make it explicit.

Factors before any elimination

For each i :

- $P(X_i \mid Z)$ is a factor on (X_i, Z) .
- $P(Y_i \mid X_i)$ is a factor on (Y_i, X_i) , but since Y_i is observed to be y_i we **restrict** this factor to that value:

$$\phi_{y_i}(X_i) = P(Y_i = y_i \mid X_i)$$

so the factor becomes a function of X_i **only** (no Y_i variable remains).

So the active factors are:

- $P(Z)$
- $\{P(X_i \mid Z)\}_{i=1}^n$
- $\{\phi_{y_i}(X_i)\}_{i=1}^n$

Eliminate X_1 (in ordering 2)

We multiply the factors that mention X_1 : $P(X_1 \mid Z)$ and $\phi_{y_1}(X_1)$. Both are functions of X_1 (the first also of Z). Sum out X_1 :

$$f_1(Z) = \sum_{x_1} P(x_1 \mid Z) \phi_{y_1}(x_1) = \sum_{x_1} P(x_1 \mid Z) P(y_1 \mid x_1).$$

Crucially: f_1 is a function of Z **only**. Y_1 does not remain a variable because it was instantiated to the observed value y_1 before elimination. So the resulting factor's scope is $\{Z\}$ (size k , if each variable has domain size k) – *not* $\{Z, Y_1\}$.

The same happens for each X_i with $i = 1, \dots, n - 1$: eliminating X_i produces $f_i(Z)$.

Multiply the $f_i(Z)$ and eliminate Z

After removing $X_1, \dots, X_{n-1}$ we have factors:

$$P(Z), P(X_n \mid Z), \text{ and } f_1(Z), \dots, f_{n-1}(Z), \phi_{y_n}(X_n).$$

Multiply the Z -only factors into one function $g(Z) = P(Z) \prod_{i=1}^{n-1} f_i(Z)$. Eliminating Z to get a factor on X_n :

$$h(X_n) = \sum_z g(z) P(X_n \mid z).$$

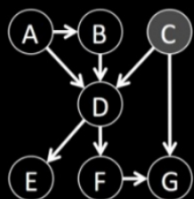
This is a function of X_n only (size k). All intermediate factors were either unary on Z or unary on the X 's – so **the largest factor size is k** .

Summary / short answer

- Your proposed $f_1(Z, Y_1)$ would be correct **only if** Y_1 were not observed.
- But because Y_1 is observed and replaced by the constant y_1 , the factor $P(Y_1 \mid X_1)$ reduces to a function of X_1 only, and after summing out X_1 we get a **Z -only** factor $f_1(Z)$.
- Therefore in ordering $X_1, \dots, X_{n-1}, Z$ the maximum factor size is k , not k^2 or k^n .

Practice

[9 pts] For the Bayes' net below, we are given the query $P(A, E \mid +c)$. All variables have binary domains. Assume we run variable elimination to compute the answer to this query, with the following variable elimination ordering: B, D, G, F .



Complete the following description of the factors generated in this process:

After inserting evidence, we have the following factors to start out with:

$$P(A), P(B|A), P(+c), P(D|A, B, +c), P(E|D), P(F|D), P(G|+c, F)$$

When eliminating B we generate a new factor f_1 as follows:

$$f_1(A, +c, D) = \sum_b P(b|A) P(D|A, b, +c)$$

This leaves us with the factors:

$$P(A), P(+c), P(E|D), P(F|D), P(G|+c, F), f_1(A, +c, D)$$

When eliminating D we generate a new factor f_2 as follows:

$$f_2(A, +c, E, F) = \sum_d P(E|d) P(F|d) f_1(A, +c, d)$$

This leaves us with the factors:

$$P(A), P(+c), P(G|+c, F), f_2(A, +c, E, F)$$

When eliminating G we generate a new factor f_3 as follows:

$$f_3(+c, F) = \sum_g P(g|+c, F)$$

This leaves us with the factors:

$$P(A), P(+c), f_2(A, +c, E, F), f_3(+c, F)$$

When eliminating F we generate a new factor f_4 as follows:

$$f_4(A, +c, E) = \sum_f f_2(A, +c, E, f) f_3(+c, f)$$

This leaves us with the factors:

$$P(A), P(+c), f_4(A, +c, E)$$

(b) [2 pts] Write a formula to compute $P(A, E \mid +c)$ from the remaining factors.

$$P(A, E \mid +c) = \frac{P(A)P(+c)f_4(A, +c, E)}{\sum_{a,e} P(a)P(+c)f_4(a, +c, e)} \text{ or alternatively: } P(A, E \mid +c) \propto P(A)P(+c)f_4(A, +c, E) \text{ and include statement that says renormalization is needed to obtain } P(A, E \mid +c).$$

(c) [2 pts] Among f_1, f_2, f_3, f_4 , which is the largest factor generated, and how large is it? Assume all variables have binary domains and measure the size of each factor by the number of rows in the table that would represent the factor.

$f_2(A, +c, E, F)$ is the largest factor generated. It has 3 non-instantiated variables, hence $2^3 = 8$ entries.

$$P(A, B, C, D, E, F, G) =$$

$$P(A) \cdot P(C) \cdot P(B|A) \cdot P(D|A, B, C) \cdot P(E|D) \cdot P(F|D) \cdot P(G|+c, F)$$

eliminating B

$$\sum_B P(B|A) \cdot P(D|A, B, C) = f_1(A, D, C=+c) \quad 2^2=4$$

eliminating D

$$\sum_D f_1(A, D, C=+c) \cdot P(E|D) \cdot P(F|D) = f_2(A, E, F, C=+c) \quad 2^3=8$$

eliminating G

$$\sum_G P(G|F, C=+c) = f_3(F, C=+c) \quad 2$$

eliminating F

$$\sum_F f_2(A, E, F, C=+c) f_3(F, C=+c) = f_4(A, E, C=+c) \quad 2^2=4$$

Even though we know for a fact that $\sum_n P(n|F, C=+c) = 1 \quad \forall F$,

still we need to consider this as a factor $f_3(F, C=+c)$ instead as a constant

since $f_3(F, C=+c) =$

$F=0, C=+c$	1
$F=1, C=+c$	1

$$P(n|F, C=+c)$$

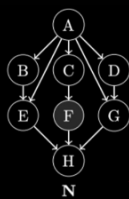
	$n=0$	$n=1$
$F=0, C=0$	m	$1-m$
$F=0, C=1$	n	$1-n$
$F=1, C=0$	o	$1-o$
$F=1, C=1$	p	$1-p$

This is what happens

instead of writing the answer as a constant, we need to write it as a factor with variable F , even if all the internal probabilities turn out to be constant.

Variable Elimination. Carry out variable elimination inference for $P(H|+f)$ in the Bayes' net N (shown to the right). Eliminate variables in the order prescribed below. All variables in this question are binary.

Answer format. To write factors made during elimination, include the factor number, arguments, elimination sum, and joined factors like so: $f_1(X, +y) = \sum_z p(z|+y)p(X|z)$.



- (a) [1 pt] What factors do we start with after incorporating evidence?
 $p(A), p(B|A), p(C|A), p(D|A), p(E|A, B), p(+f|C), p(G|A, D), p(H|E, +f, G)$
- (b) [1 pt] Eliminate A to make a new factor f_1 .
 $f_1(B, C, D, E, G) = \sum_a p(a)p(B|a)p(C|a)p(D|a)p(E|a, B)p(G|a, D)$
- (c) [1 pt] Eliminate B to make a new factor f_2 .
 $f_2(C, D, E, G) = \sum_b f_1(b, C, D, E, G)$
- (d) [1 pt] Eliminate C to make a new factor f_3 .
 $f_3(D, E, +f, G) = \sum_c f_2(c, D, E, G)p(+f|c)$
- (e) [1 pt] Eliminate D to make a new factor f_4 .
 $f_4(+f, E, G) = \sum_d f_3(d, E, G, +f)$
- (f) [1 pt] Eliminate E to make a new factor f_5 .
 $f_5(+f, G, H) = \sum_e f_4(e, G, +f)p(H|e, +f, G)$
- (g) [1 pt] Eliminate G to make a new factor f_6 .
 $f_6(+f, H) = \sum_g f_5(+f, g, H)$
- (h) [1 pt] Briefly explain how to compute $p(+h|+f)$ from the set of factors returned by variable elimination.
 Normalize: $p(+h|+f) = \frac{f_6(+f, +h)}{f_6(+f, +h) + f_6(+f, -h)}$
- (i) [1 pt] Which factor generated during variable elimination has the largest number of unobserved variables?
 $f_1(B, C, D, E, G)$
 How many table entries are needed to represent it? (Recall all variables are binary.) 2^5

$$P(A, B, C, D, E, F, G, H) = P(A) \cdot P(B|A) \cdot P(C|A) \cdot P(D|A) \cdot P(E|B, A) \cdot P(F|C, A) \cdot P(G|A, D) \cdot P(H|E, G, +f)$$

$$\sum_A = P(A) \cdot P(B|A) \cdot P(C|A) \cdot P(D|A) \cdot P(E|B, A) \cdot P(F|C, A)$$

$$= f_1(B, C, D, E, G) \quad (2^5 = 32 \text{ variables})$$

$$\sum_B = f_1(B, C, D, E, G) = f_2(C, D, E, G) \quad (2^4 = 16)$$

$$\sum_C = f_2(C, D, E, G) \cdot P(+f|C) = f_3(D, E, G, +f) \quad (2^3 = 8)$$

$$\sum_D = f_3(D, E, G, +f) = f_4(E, G, +f) \quad (2^2 = 4)$$

$$\sum_E = f_4(E, G, +f) \cdot P(H|E, G, +f) = f_5(G, H, +f) \quad (2^2 = 4)$$

$$\sum_G = f_5(G, H, +f) = f_6(H, +f) \quad (2^1 = 2)$$

$$P(H|+f) = \frac{f_6(H, +f)}{\sum_H f_6(H, +f)}$$

$$f_1(B, C, D, E, G) \rightarrow 2^5 = 32 \text{ entries in factor table}$$

Approximate Inference

Exact inference: $P(X | E) = 0.71828$

- Get exact probability value
- Using the inference steps derived by probabilistic formula
- Need exponential time complexity

Approximate inference: $P(X | E) \approx 0.71$

- Get approximate probability value
- Using sampling theorem
- Need only polynomial time complexity, fast computation

Inference: getting a sample is faster than computing the right answer (e.g. with variable elimination)

- Estimate some values by random number generation

Sampling
 $\Rightarrow$ Random number generation

- Draw N samples from a known distribution P
- Generate N random numbers from a known distribution S

Estimation

- Compute an approximate probability $\hat{P}$, which approximates the real posterior probability $P(X | E)$

Recap - Bayes Theorem

The Posterior

The distribution of θ after observing the data

The Prior

The distribution of θ before observing the data

$$P(\theta|X) \propto P(X|\theta) \cdot P(\theta)$$

The Likelihood

The distribution of the data conditional on θ

How to take samples from an arbitrary given distribution?

- Step 1: Get sample u from uniform distribution over $[0, 1]$
 E.g. `random()` in python
- Step 2: Convert this sample u into an outcome for the given distribution by having each outcome associated with a sub-interval of $[0, 1]$ with sub-interval size equal to probability of the outcome

eg.

C	P(C)
red	0.6
green	0.1
blue	0.3

$0 \leq u < 0.6, \rightarrow C = \text{red}$

$0.6 \leq u < 0.7, \rightarrow C = \text{green}$

$0.7 \leq u < 1, \rightarrow C = \text{blue}$

- If `random()` returns $u = 0.83$, then our sample is $C = \text{blue}$
- E.g., after sampling 8 times:

Types of Sampling in Bayes Net

Prior Sampling

Rejection Sampling

Likelihood Weighting

Gibbs Sampling

Prior Sampling

① start with C let u, v, w, x be $\text{unif}(0,1)$ iid RVs

$$P(C)$$

+c	0.5
-c	0.5

$C = \begin{cases} +c, & 0 \leq u < 0.5 \\ -c, & 0.5 \leq u \leq 1 \end{cases}$ Sample ① $u = 0.2 \Rightarrow +c$

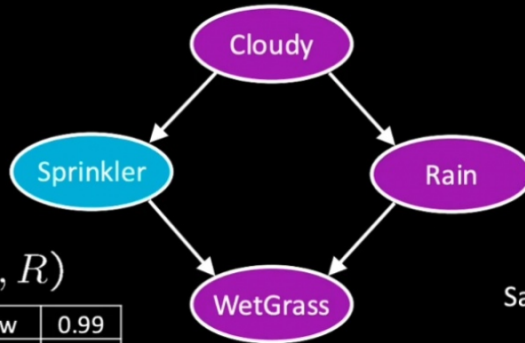
② Since sample 1 $c = +c$ we only look at $P(S|c=+c)$

$P(S|C)$

+c	+s	0.1
	-s	0.9
-c	+s	0.5
	-s	0.5

$S = \begin{cases} +s, & 0 \leq v \leq 0.1 \\ -s, & 0.1 < v < 1 \end{cases}$

Sample ② $v = 0.4 \Rightarrow S = -s$



$P(R|C)$

+c	+r	0.8
	-r	0.2
-c	+r	0.2
	-r	0.8

③ Since sample 1, $C = +c$ we look at $P(R|c=+c)$

$R = \begin{cases} +r, & 0 \leq w \leq 0.8 \\ -r, & 0.8 < w < 1 \end{cases}$

Sample ③ $w = 0.55 \Rightarrow R = +r$

$P(W|S, R)$

+s	+r	+w	0.99
		-w	0.01
	-r	+w	0.90
		-w	0.10
-s	+r	+w	0.90
		-w	0.10
	-r	+w	0.01
		-w	0.99

④ we look at $P(W|S=-s, R=+r)$

$W = \begin{cases} +w, & 0 \leq x \leq 0.9 \\ -w, & 0.9 < x < 1 \end{cases}$

Sample ④ $x = 0.8 \Rightarrow W = +w$

Samples:

+c, -s, +r, +w
-c, +s, -r, +w
...



- For $i=1, 2, \dots, n$
 - Sample x_i from $P(X_i | \text{Parents}(X_i))$
- Return $(x_1, x_2, \dots, x_n)$

→ Similarly

We'll get a bunch of samples from the BN:

- +c, -s, +r, +w
- +c, +s, +r, +w
- c, +s, +r, -w
- +c, -s, +r, +w
- c, -s, -r, +w



If we want to know $P(W)$

- We have counts $\langle +w:4, -w:1 \rangle$
- Normalize to get $P(W) = \langle +w:0.8, -w:0.2 \rangle$
- This will get closer to the true distribution with more samples
- Can estimate anything else, too
- What about $P(C|+w)$? $P(C|+r, +w)$? $P(C|-r, -w)$?
- Fast: can use fewer samples if less time (what's the drawback?)

Practice

Prior Sampling: What does the value $\frac{N(+a, -b, +c)}{N}$ approximate?

- A. $P(+a, -b, +c)$
- B. $P(+c | +a, -b)$
- C. $P(+c | -b)$
- D. $P(+c)$
- E. I don't know



$P(A)$

+a	1/2
-a	1/2

$P(B|A)$

+a	+b	1/10
-a	+b	9/10
-a	-b	1/2
-a	-b	1/2

$P(C|B)$

+b	+c	4/5
-b	+c	1/5
-b	-c	1
-b	-c	0

b) $P(A, B, C) = P(A) \cdot P(B|A) \cdot P(C|B)$

$P(+a, -b, -c) = 1/2 \cdot 9/10 \cdot 1 = \frac{9}{20}$

$\Rightarrow$ We can expect $\frac{9}{20} \cdot 1000 = 450$ samples

c) $P(-a, +b, -c) = 1/2 \cdot 1/2 \cdot 1/5 = \frac{1}{20}$

$= \frac{1000}{20} = 50$ samples

- b) how many can we expect given we took 1000 samples
- c) how many samples of $(-a, +b, -c)$ // // // //

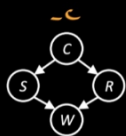
Rejection Sampling

Let's say we want $P(C)$

- No point keeping all samples around
- Just tally counts of C as we go

Let's say we want $P(C|+s)$

- Same thing: tally C outcomes, but ignore (reject) samples which don't have $S=+s$
- This is called rejection sampling
- It is also consistent for conditional probabilities (i.e., correct in the limit)



+C, -S, +R, +W
+C, +S, +R, +W
-C, +S, +R, -W
+C, -S, +R, +W
-C, -S, -R, +W

- Works over prior sampling
- Easier to calculate marginals
- Possible to calculate conditional

- Given $P(Q|evidence)$
- Reject samples that don't follow evidence
- Tally observations of P
- Approximation = $\frac{\text{tally}(P)}{\# \text{unrejected samples}} = \frac{N(P)}{N(evidence)}$

Algorithm

IN: evidence instantiation

For $i=1, 2, \dots, n$

- Sample x_i from $P(X_i | \text{Parents}(X_i))$
- If x_i not consistent with evidence
 - Reject: Return, and no sample is generated in this cycle

Return $(x_1, x_2, \dots, x_n)$

E.g., estimate $P(\text{Rain} | \text{Sprinkler} = \text{true})$ using 100 samples

27 samples have $\text{Sprinkler} = \text{true}$

Of these, 8 have $\text{Rain} = \text{true}$ and 19 have $\text{Rain} = \text{false}$.

$\hat{P}(\text{Rain} | \text{Sprinkler} = \text{true}) = \text{NORMALIZE}(\{8, 19\}) = \{0.296, 0.704\}$

Similar to a basic real-world empirical estimation procedure

Practice

Assume the following Bayes' net, and the corresponding distributions over the variables in the Bayes' net:

$P(A)$	$P(B A)$	$P(C B)$	$P(D C)$
-a 3/4	-a -b 2/3	-b -c 1/4	-c -d 1/8
+a 1/4	-a +b 1/3	-b +c 3/4	-c +d 7/8
	+a -b 4/5	+b -c 1/2	+c -d 5/6
	+a +b 1/5	+b +c 1/2	+c +d 1/6

(a) You are given the following samples:

+a	+b	-c	-d
+a	-b	+c	-d
-a	+b	+c	-d
-a	-b	+c	-d

(i) Assume that these samples came from performing Prior Sampling, and calculate the sample estimate of $P(+c)$.

5/8

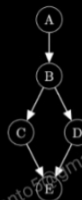
(ii) Now we will estimate $P(+c | +a, -d)$. Above, clearly cross out the samples that would not be used when doing Rejection Sampling for this task, and write down the sample estimate of $P(+c | +a, -d)$ below.

2/3

$$a) P(C=+c) = \frac{N(+c)}{N} = 5/8$$

$$b) P(+c | +a, -d) = \frac{N(+c)}{N(+a, -d)} = 2/3$$

Assume we are given the following Bayes' net, with the associated conditional probability tables (CPTs).



A	$P(A)$
+a	0.5
-a	0.5

A	B	$P(B A)$
+a	+b	0.2
+a	-b	0.8
-a	+b	0.5
-a	-b	0.5

B	C	$P(C B)$
+b	+c	0.4
+b	-c	0.6
-b	+c	0.8
-b	-c	0.2

B	D	$P(D B)$
+b	+d	0.2
+b	-d	0.8
-b	+d	0.2
-b	-d	0.8

C	D	E	$P(E C, D)$
+c	+d	+e	0.6
+c	+d	-e	0.4
+c	-d	+e	0.2
+c	-d	-e	0.8
-c	+d	+e	0.4
-c	+d	-e	0.6
-c	-d	+e	0.8
-c	-d	-e	0.2

You are given a set of the following samples, but are not told whether they were collected with rejection sampling or likelihood weighting.

-a	-b	+c	+d	+e
-a	+b	+c	-d	+e
-a	-b	-c	-d	+e
-a	-b	+c	-d	+e
-a	+b	+c	+d	+e

Throughout this problem, you may answer as either numeric expressions (e.g. $0.1 \cdot 0.5$) or numeric values (e.g. 0.05).

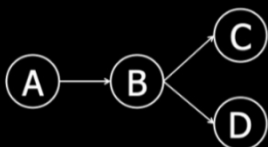
(a) [2 pts] Assuming these samples were generated from rejection sampling, what is the sample based estimate of $P(+b | -a, +e)$?

$$= \frac{N(+b)}{N(+e)} = 2/5 = 0.4$$

Answer: 0.4

The answer is the number of samples satisfying the query variable's assignment (in this case, $B = +b$ divided by the total number of samples, so the answer is $2 / 5 = 0.4$).

Assume the following Bayes net, and the corresponding distributions over the variables in the Bayes net:



A	$P(A)$
+a	1/5
-a	4/5

B	C	$P(C B)$
+b	+c	1/4
+b	-c	3/4
-b	+c	2/5
-b	-c	3/5

A	B	$P(B A)$
+a	+b	1/5
+a	-b	4/5
-a	+b	1/2
-a	-b	1/2

B	D	$P(D B)$
+b	+d	1/2
+b	-d	1/2
-b	+d	4/5
-b	-d	1/5

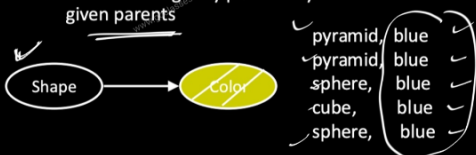
(a) [2 pts] Your task is now to estimate $P(+b, -a, -c, -d)$ using rejection sampling. Below are some samples that have been produced by prior sampling (that is, the rejection stage in rejection sampling hasn't happened yet). Cross out the samples that would be rejected by rejection sampling:

~~-a -b +c +d~~ -a -b -c -d
~~+a -b -c +d~~ -a +b +c +d
~~-a -b -c -d~~ +a -b -c -d

(b) [1 pt] Using those samples, what value would you estimate for $P(+b, -a, -c, -d)$ using rejection sampling? $\Rightarrow 0$

Likelihood Weighting

- Idea: fix evidence variables and sample the rest
- Problem: sample distribution not consistent!
- Solution: weight by probability of evidence given parents



- IN: evidence instantiation
- $w = 1.0$
- for $i=1, 2, \dots, n$
 - if X_i is an evidence variable
 - $X_i = \text{observation } x_i \text{ for } X_i$
 - Set $w = w * P(x_i | \text{Parents}(X_i))$
 - else
 - Sample x_i from $P(X_i | \text{Parents}(X_i))$
- return $(x_1, x_2, \dots, x_n), w$

- Sampling distribution if z sampled and e fixed evidence

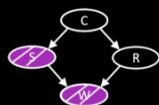
$$S_{WS}(z, e) = \prod_{i=1}^l P(z_i | \text{Parents}(z_i))$$

- Now, samples have weights

$$w(z, e) = \prod_{i=1}^m P(e_i | \text{Parents}(e_i))$$

- Together, weighted sampling distribution is consistent

$$S_{WS}(z, e) \cdot w(z, e) = \prod_{i=1}^l P(z_i | \text{Parents}(z_i)) \prod_{i=1}^m P(e_i | \text{Parents}(e_i)) = P(z, e)$$



eg. $A \rightarrow B$

A	$P(A)$
0	0.5
1	0.5

	$P(B A)$	B=0	B=1
A=0	1	0	1
A=1	0	1	1

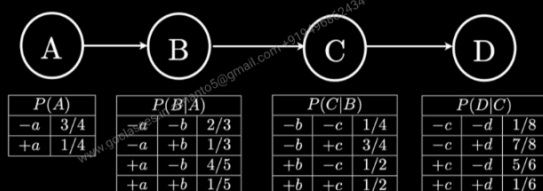
a) without weight $\rightarrow$ samples = $\{A=1, B=1\}$ 1/2 samples
 $\{A=0, B=1\}$ 1/2 samples

from sample $P(A|B) = 0.5$ wrong.

b) with weighting $\{A=1, B=1\} \times P(B=1|A=1) = 1$
 $\{A=0, B=1\} \times P(B=1|A=0) = 0$

$$P(A=1, B=1) = \frac{\frac{N}{2} \times 1}{\frac{N}{2} \times 1 + \frac{N}{2} \times 0} = 1$$

Assume the following Bayes' net, and the corresponding distributions over the variables in the



You are given the following samples—*for prior sampling*

+a	+b	-c	-d	+a	-b	-c	+d
+a	-b	+c	-d	+a	+b	+c	-d
-a	+b	+c	-d	-a	+b	-c	+d
-a	-b	+c	-d	-a	-b	+c	-d

Using Likelihood Weighting Sampling to estimate $P(-a | +b, -d)$, the following samples were obtained. Fill in the weight of each sample in the corresponding row.

These are samples for Likelihood Sampling

Sample

$$\begin{aligned}
 -a \quad +b \quad +c \quad -d &\rightarrow P(-d|+c) \cdot P(+b|-a) = 5/6 \cdot 1/3 = 5/18 \\
 +a \quad +b \quad +c \quad -d &\rightarrow P(-d|+c) \cdot P(+b|+a) = 5/6 \cdot 1/5 = 1/6 \\
 +a \quad +b \quad -c \quad -d &\rightarrow P(-d|-c) \cdot P(+b|+a) = 1/8 \cdot 1/5 = 1/40 \\
 -a \quad +b \quad -c \quad -d &\rightarrow P(-d|-c) \cdot P(+b|-a) = 1/8 \cdot 1/3 = 1/24
 \end{aligned}$$

From the weighted samples in the previous question, estimate $P(-a | +b, -d)$.

$$P(-a | +b, -d) = \frac{5/18 + 1/24}{\frac{5}{18} + \frac{1}{6} + \frac{1}{40} + \frac{1}{24}}$$

(6 points) The Sound of (Silent) Music

You are an exobiologist, studying the wide range of life in the universe. You are also an avid dancer and have an excellent model of the way species invent dancing. The key variables are:

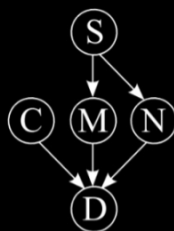
Sound sensing (S): Whether or not a species has the ability to sense sound

Cold climate (C): Whether or not the native planet of the species has cold weather

Music (M): Whether or not the species invented music

Non-verbal communication (N): Whether or not the species has any form of non-verbal communication

You model the relationships between these variables and dancing (D) using the Bayes net specified to the right.



You want to know how likely it is for a dancing, sound-sensing species to invent music, according to this Bayes net. However, you're a doctor, not a dynamic programmer, and you can't bear the thought of variable elimination this late in the exam. So, you decide to do inference via sampling.

You use prior sampling to draw the samples below:

S	C	M	N	D
-s	+c	+m	-n	+d
+s	+c	-m	-n	-d
+s	-c	-m	+n	-d
+s	+c	+m	-n	+d
+s	+c	-m	+n	+d
+s	-c	-m	+n	-d
+s	-c	-m	-n	-d
+s	+c	+m	+n	+d
+s	+c	-m	+n	-d
-s	-c	-m	-n	-d

$$P(-m | +d, +s) = \dots$$

(a) (2 pt) Based on rejection sampling using the samples above, what is the answer to your query, $P(-m | +d, +s)$?

b) Likelihood weighting $P(+m | -s, -d)$

$$= \frac{0.1 + 0.1}{0.1 + 0.1 + 0.8 + 0.4} = \frac{0.2}{1.4} = \frac{1}{7}$$

g) Rejection Sampling

$$P(-m | +d, +s) = 1/3$$

While your sampling method has worked fairly well in many cases, for rare cases (like species that can't sense sound) your results are less accurate as rejection sampling rejects almost all of the data. You decide to use likelihood weighting instead. The conditional probabilities of the Bayes net are listed below.

C	M	N	D	$P(D C, M, N)$
+c	+m	+n	+d	0.9
+c	+m	+n	-d	0.1
+c	+m	-n	+d	0.8
+c	+m	-n	-d	0.2
+c	-m	+n	+d	0.8
+c	-m	+n	-d	0.2
+c	-m	-n	+d	0.2
+c	-m	-n	-d	0.8
-c	+m	+n	+d	0.8
-c	+m	+n	-d	0.2
-c	+m	-n	+d	0.5
-c	+m	-n	-d	0.5
-c	-m	+n	+d	0.6
-c	-m	+n	-d	0.4
-c	-m	-n	+d	0.1
-c	-m	-n	-d	0.9

S	M	$P(M S)$
+s	+m	0.8
+s	-m	0.2
-s	+m	0.1
-s	-m	0.9

S	$P(S)$
+s	0.9
-s	0.1

S	N	$P(N S)$
+s	+n	0.7
+s	-n	0.3
-s	+n	0.9
-s	-n	0.1

C	$P(C)$
+c	0.5
-c	0.5

You now wish to compute the probability that a species that has no sound-sensing (-s) or dancing (-d) nonetheless has music (+m), using likelihood weighting. I.e., you want $P(+m | -s, -d)$.

(b) (2 pt) You draw the samples below, using likelihood weighting. For each of these samples, indicate its weight.

S	C	M	N	D	weight
-s	+c	+m	+n	-d	0.01
-s	+c	-m	-n	-d	0.08
-s	-c	-m	+n	-d	0.24
-s	+c	+m	+n	-d	0.01

$$\begin{aligned}
 &\rightarrow P(-d | +c, +m, +n) \cdot P(-s) \\
 &\rightarrow P(-d | +c, -m, -n) \cdot P(-s) \\
 &\rightarrow P(-d | -c, -m, +n) \cdot P(-s) \\
 &\rightarrow P(-d | +c, +m, +n) \cdot P(-s)
 \end{aligned}$$



Likelihood-weighted sampling suffers from **weight degeneracy** (most samples get extremely small weights when evidence is deep in the network), leading to high variance and ineffective sampling. **Gibbs sampling** solves this by not using weights at all—it samples directly from the posterior distribution, keeping the evidence fixed and resampling only the non-evidence variables, which avoids weight collapse.

Detailed explanation

Problem with Likelihood-Weighted Sampling (LWS)

LWS works by:

- Fixing the evidence variables to their observed values
- Sampling all other variables forward in topological order
- Assigning a weight to each sample equal to the likelihood of the evidence given the ancestors

But the issue:

If the evidence variable is deep in the network (many ancestors), the likelihood weight becomes:

- Extremely small for most samples
- Highly concentrated on a few samples

This is called **weight degeneracy** or **weight collapse**.

As a result:

- Effective sample size becomes very small
- Variance becomes extremely high
- You need enormous sample counts to get stable estimates

This makes LWS perform poorly when the evidence is low-probability or far down the graph.

Gibbs Sampling

Procedure: keep track of a full instantiation $x_1, x_2, \dots, x_n$. Start with an arbitrary instantiation consistent with the evidence. Sample one variable at a time, conditioned on all the rest, but keep evidence fixed. Keep repeating this for a long time.

Property: in the limit of repeating this infinitely many times the resulting sample is coming from the correct distribution

Rationale: both upstream and downstream variables condition on evidence.

In contrast: likelihood weighting only conditions on upstream evidence, and hence weights obtained in likelihood weighting can sometimes be very small. Sum of weights over all samples is indicative of how many "effective" samples were obtained, so want high weight.

This happens for samples with upstream variable values that are very unlikely given evidence
 $\Rightarrow P(\text{evidence} | \text{parents})$ becomes low
 $\Rightarrow$ weights very low.
 Example: $A \rightarrow B$ please page

→ In Likelihood weighted sampling what is meant by

✓ Why evidence influences downstream variables

"Downstream" = child variables.

Because sampling proceeds **forward**, once an evidence variable is fixed to some value:

- its children will be sampled from $P(\text{child} \mid \text{evidence})$

So the evidence affects the distribution of all the nodes that come after it.

✓ Evidence influences its descendants (downstream)

✗ Because they are conditioned on the evidence in the CPT.

✗ Why evidence does not influence upstream variables

"Upstream" = parent variables.

In LWS:

- Parents are sampled **before** evidence is seen.
- They are **not conditioned on the evidence**, because the sampling moves only forward.
- Instead, the evidence only contributes a weight after parents have already been sampled.

Thus:

- The "parent" values that were sampled might be completely inconsistent with the evidence.
- They still remain in the sample; only the *weight* becomes tiny.

This is the root problem of LWS:

samples inconsistent with evidence are not resampled—they are just down-weighted, often to zero.

✓ Evidence **does NOT** influence ancestors (upstream)

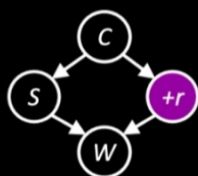
✗ Because LWS never "goes backwards" to update parents using Bayes' rule.

Gibbs

Gibbs Sampling Example: $P(S \mid +r)$

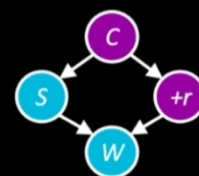
Step 1: Fix evidence

- $R = +r$



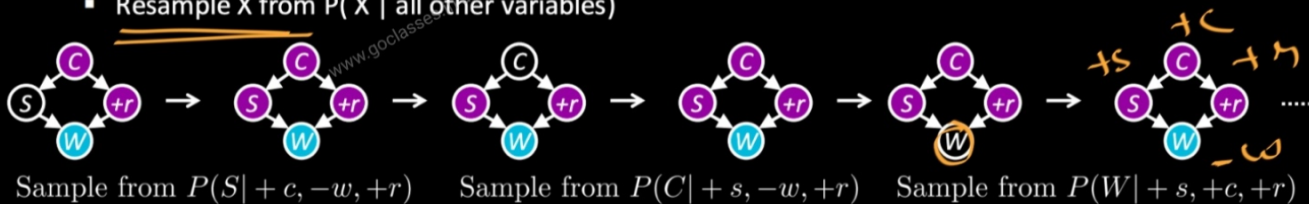
Step 2: Initialize other variables

- Randomly



Steps 3: Repeat

- Choose a non-evidence variable X
- Resample X from $P(X \mid \text{all other variables})$



Sample from $P(S \mid +c, +r, -w)$

$$\begin{aligned} P(S \mid +c, +r, -w) &= \frac{P(S, +c, +r, -w)}{P(+c, +r, -w)} \\ &= \frac{P(S, +c, +r, -w)}{\sum_s P(s, +c, +r, -w)} \\ &= \frac{P(+c)P(S|+c)P(+r|+c)P(-w|S, +r)}{\sum_s P(+c)P(s|+c)P(+r|+c)P(-w|s, +r)} \\ &= \frac{P(+c)P(S|+c)P(+r|+c)P(-w|S, +r)}{\sum_s P(+c)P(s|+c)P(+r|+c)P(-w|s, +r)} \\ &= \frac{P(+c)P(S|+c)P(+r|+c)P(-w|S, +r)}{\sum_s P(s|+c)P(-w|s, +r)} \end{aligned}$$

→ $\sum_s$ taken inside
→ All CPTs w/o S get cancelled

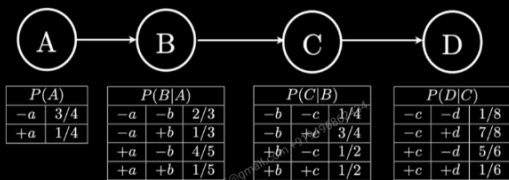
Many things cancel out – only CPTs with S remain!

More generally: only CPTs that have resampled variable need to be considered, and joined together



→ CPT with resampled variables
→ Join together to get new CPT
 $P(\text{resampled variable} \mid \text{all other variables})$

Assume the following Bayes' net, and the corresponding distributions over the variables in the Bayes' net:



You are given the following samples:

+a	+b	-c	-d	+a	-b	-c	+d
+a	-b	+c	-d	+a	+b	+c	-d
-a	+b	+c	-d	-a	+b	-c	+d
-a	-b	+c	-d	-a	-b	+c	-d

evidence is upstream variable

Which query is better suited for likelihood weighting, $P(D | A)$ or $P(A | D)$? Justify your answer in one sentence.

$P(D | A)$ is better suited for likelihood weighting sampling, because likelihood weighting conditions only on upstream evidence.

Recall that during Gibbs Sampling, samples are generated through an iterative process.

Assume that the only evidence that is available is $A = +a$. Clearly fill in the circle(s) of the sequence(s) below that could have been generated by Gibbs Sampling.

Sequence 1

1:	+a	-b	-c	+d
2:	+a	-b	-c	+d
3:	+a	-b	+c	+d

X

Sequence 2

1:	+a	-b	-c	+d
2:	+a	-b	-c	-d
3:	-a	-b	-c	+d

Sequence 3

1:	+a	-b	-c	+d
2:	+a	-b	-c	-d
3:	+a	+b	-c	-d

Sequence 4

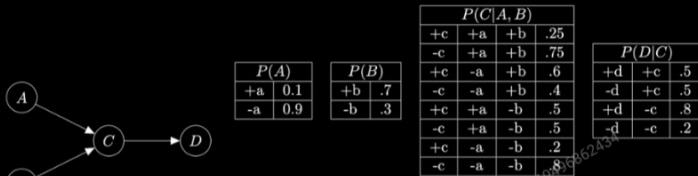
1:	+a	-b	-c	+d
2:	+a	-b	-c	-d
3:	+a	+b	-c	+d

Both variable b & d changed → Not possible

Gibbs sampling updates one variable at a time and never changes the evidence.

The first and third sequences have at most one variable change per row, and hence could have been generated from Gibbs sampling. In sequence 2, the evidence variable is changed. In sequence 4, the second and third samples have both B and D changing.

Assume you are given the following Bayes' net and the corresponding distributions over the variables in the Bayes' net.



(b) [?? pts] Next, assume we observed both $A = +a$ and $D = +d$. What are the weights for the following samples under likelihood weighting sampling?

Sample	Weight
$(+a, -b, +c, +d)$	$P(+a) \cdot P(+d +c) = 0.1 * 0.5 = 0.05$
$(+a, -b, -c, +d)$	$P(+a) \cdot P(+d -c) = 0.1 * 0.8 = 0.08$
$(+a, +b, -c, +d)$	$P(+a) \cdot P(+d -c) = 0.1 * 0.8 = 0.08$

(c) [?? pts] Given the samples in the previous question, estimate $P(-b | +a, +d)$.

$$p(-b | +a, +d) = \frac{0.05 + 0.08}{0.05 + 0.08 + 0.08} = \frac{13}{21}$$

(d) [?? pts] Assume we need to (approximately) answer two different inference queries for this graph: $P(C | +a)$ and $P(C | +d)$. You are required to answer one query using likelihood weighting and one query using Gibbs sampling. In each case you can only collect a relatively small amount of samples, so for maximal accuracy you need to make sure you cleverly assign algorithm to query based on how well the algorithm fits the query. Which query would you answer with each algorithm?

Algorithm	Query	Algorithm	Query
Likelihood Weighting	$P(C +a)$	Gibbs Sampling	$P(C +d)$

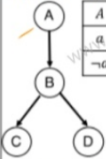
Justify your answer:

You should use Gibbs sampling to find the query answer $P(C | +d)$. This is because likelihood weighting only takes upstream evidence into account when sampling. Therefore, Gibbs, which utilizes both upstream and downstream evidence, is more suited to the query $P(C | +d)$ which has downstream evidence.

Q.32

Consider the Bayes Net containing four Boolean random variables (A, B, C, D), with the following convention: $A = \text{True} \Rightarrow A = a$, and $A = \text{False} \Rightarrow A = \neg a$; and similarly for the other variables. The conditional probability tables for the nodes in the network are also indicated in the figure. The following samples were generated through likelihood weighting:

$s_1: (\neg a, \neg b, \neg c, \neg d); s_2: (\neg a, b, \neg c, \neg d); s_3: (\neg a, \neg b, \neg c, d); s_4: (\neg a, b, \neg c, d)$



A	P(A)
a	1/5
¬a	4/5

A	B	P(B A)
a	b	1/5
¬a	b	1/2
a	¬b	4/5
¬a	¬b	1/2

B	C	P(C B)
b	c	1/4
¬b	c	3/4
b	¬c	2/5
¬b	¬c	3/5

B	D	P(D B)
b	d	1/2
¬b	d	4/5
b	¬d	1/2
¬b	¬d	1/5

Estimate the likelihood weight of each sample and thereby estimate $P(b|\neg a, \neg c)$

(A) $s_1: 0.48, s_2: 0.32, s_3: 0.48, s_4: 0.32, P(b|\neg a, \neg c) = 0.4$

(B) $s_1: 0.48, s_2: 0.32, s_3: 0.48, s_4: 0.32, P(b|\neg a, \neg c) = 0.64$

(C) $s_1: 0.32, s_2: 0.48, s_3: 0.48, s_4: 0.32, P(b|\neg a, \neg c) = 0.64$

(D) $s_1: 0.48, s_2: 0.32, s_3: 0.32, s_4: 0.32, P(b|\neg a, \neg c) = 0.4$

$$s_1 \rightarrow P(\bar{a}) \cdot P(\bar{c}|\bar{b}) = 4/5 \cdot 3/5 = \frac{12}{25} = 0.48$$

$$s_2 \rightarrow P(\bar{a}) \cdot P(c|\bar{b}) = 4/5 \cdot 2/5 = 8/25 = 0.32$$

$$s_3 \rightarrow P(\bar{a}) \cdot P(\bar{c}|\bar{b}) = 12/25 = 0.48$$

$$s_4 \rightarrow P(\bar{a}) \cdot P(c|\bar{b}) = 8/25 = 0.32$$

$$P(b|\bar{a}, \bar{c}) = \frac{8/15 + 8/15}{\frac{8}{15} + \frac{8}{15} + \frac{12}{15} + \frac{12}{15}} = \frac{8}{20} = 0.4$$