COMP 341 Intro to AI
Bayesian Networks – Approximate Inference

Asst. Prof. Barış Akgün

Koç University

Probabilistic Inference Methods

- Enumeration (exact, exponential complexity)
- Variable elimination (exact, worst-case exponential complexity, often better)
- Exact Inference is NP-complete (surprised?)
- Sampling(approximate)

Approximate Inference

- It takes too long and too much space to calculate the desired probabilities from big BNs

- Idea: Approximate the posterior probability by sampling from the BN
  - Sampling is very popular in a variety of fields, and it is simple
  - Inference: sampling is faster than computing the right answer
  - Learning: Sample from an unknown distribution

- Basic Idea:
  - Draw N samples from a sampling distribution S
  - Compute an approximate posterior probability
  - Show this converges to the true probability P

Brief Detour-Discrete Distributions

- **Probability Mass Function**: Probability that a discrete random variable, X, is exactly equal to some value. So far, we used probability tables to represent this.

$$f_X(x) = P(X = x)$$

$$\sum_x f_X(x) = 1$$

X	P(X)
1	0.6
2	0.1
3	0.3

- **Cumulative Distribution Function**: Probability that a random variable, X, will take a value less than* x.

$$F_X(x) = P(X = x)$$

$$P(a \leq X < b) = F_X(b) - F_X(a)$$

Returns the corresponding interval of a particular value given the CDF

X	F(X)	I(X)
1	0.6	[0,0.6)
2	0.7	[0.6,0.7)
3	1.0	[0.7,1.0)

*This is a convention. CDF could also be defined with “less than or equal to”. This distinction does not matter for continuous variables

Sampling from a Known Distribution

- **Step 1**: Get sample *u* from uniform distribution over [0, 1)
- **Step 2**: Convert this sample *u* into an outcome for the given distribution by having each outcome associated with an interval of [0,1) with the interval size equal to the probability of the outcome
- E.g. u = 0.83, what is the sample?
- Sample n times, then divide the counts by n to get the P(.)
- After sampling 10 times we get:
  - b-r-r-r-r-r-b-r-r-g
  - P(r)= 0.7, P(b)=0.2, P(g) = 0.1
- What about 100 times?
  - P(r)= 0.62, P(b)=0.27, P(g) = 0.11
- 1000 times?
  - P(r)= 0.603, P(b)=0.297, P(g) = 0.1

C	P(C)
red	0.6
green	0.1
blue	0.3

$$0 \leq u < 0.6, \rightarrow C = red$$
$$0.6 \leq u < 0.7, \rightarrow C = green$$
$$0.7 \leq u < 1, \rightarrow C = blue$$

Note that you can use the CDF to get these intervals by assuming an order even if the values are not numbers

Python: Try at Home

```
from random import random
def cumsum(f):
  total = 0
  for x in f:
    total += x
    yield total
def getVal(u, cs):
  for i in range(0,len(cs)):
    if u < cs[i]:
      return i
probs = [0.6,0.3,0.1]
numIter = 1000
cs = list(cumsum(probs))
counts = [0]*len(probs)
for i in range(0,numIter):
  #Step 1:
  u = random()
  #Step 2:
  t = getVal(u,cs)
  counts[t] += 1
print([num/float(numIter) for num in counts])
```

Play around with these to see what happens (make sure probs sum up to 1!)

Bayesian Network Recap

- Represented as a directed acyclic graph with nodes as variables
- Implicitly encode the joint probability distribution as a product of local conditional distributions

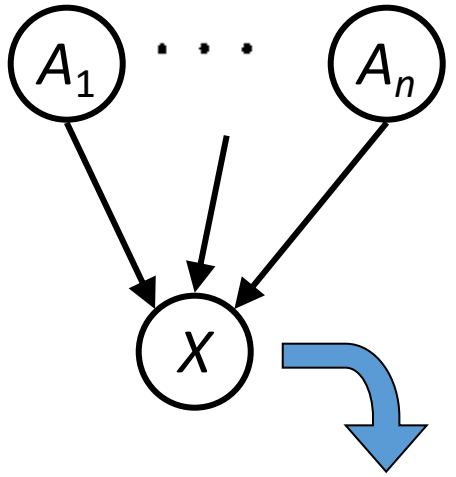
$$P(X_1, X_2, \dots, X_n) = \prod_{i=1}^{n} P\big(X_i | Parents(X_i)\big)$$

- Inference: Calculating a useful quantity from a joint probability distribution.
  - Posterior Probability:

    $$\mathrm{P}(Q | E_1 = e_1, \dots, E_k = e_k)$$
  - Most Likely Explanation:

    $$\mathrm{argmax_q}\ \mathrm{P}(Q = q | E_1 = e_1, \dots, E_k = e_k)$$

- Exact inference has exponential complexity (for generic BNs)
- Approximate Inference and Sampling



$$P(X | A_1 \dots A_n)$$

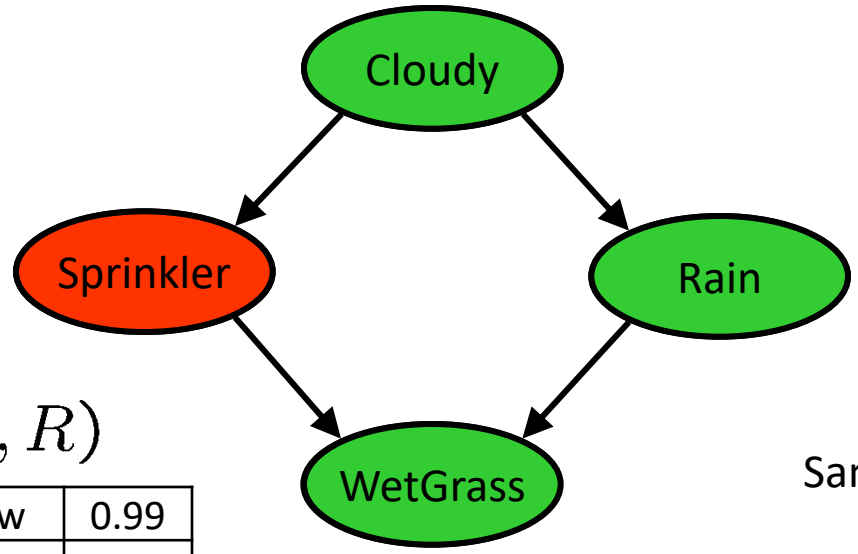
We can answer the following by now:

- Why do we want the joint distribution?
- Why are we using BNs?
- What is inference?
- What is enumeration and variable elimination?
- Why approximate inference?

Sampling in Bayes Nets

- Prior Sampling
- Rejection Sampling
- Likelihood Weighting
- Gibbs Sampling

Prior Sampling



$P(C)$

+c	0.5
-c	0.5

$P(S|C)$

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

$P(R|C)$

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

$P(W|S,R)$

+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

Samples:

+c, -s, +r, +w

-c, +s, -r, +w

…

Prior Sampling

- Given a BN, specifying $P(X_1, \ldots, X_n)$, with the condition $X_j \notin Parents(X_i)$, for $j > i$

```
for i = 1 to n do
    x_i = sample(P(X_i | Parents(X_i)))
return (x_1, ..., x_n)
```

- The sampling uses the values of $Parents(X_i)$ from obtained from previous steps

Prior Sampling

- This process generates samples with probability:

$$S_{PS}(x_1 \ldots x_n) = \prod_{i=1}^{n} P(x_i|\text{Parents}(X_i)) = P(x_1 \ldots x_n)$$

  …i.e. the BN's joint probability

- Let the number of samples of an event be $N_{PS}(x_1 \ldots x_n)$

- Then $\lim_{N\to\infty} \hat{P}(x_1, \ldots, x_n) = \lim_{N\to\infty} N_{PS}(x_1, \ldots, x_n)/N$
  $= S_{PS}(x_1, \ldots, x_n)$
  $= P(x_1 \ldots x_n)$

- I.e., the sampling procedure is consistent

Example

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

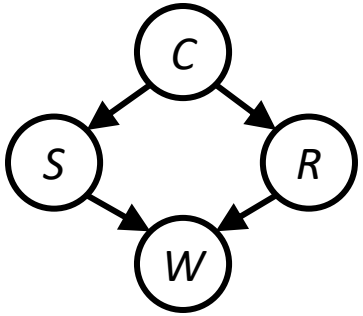
  +c, -s, +r, +w

  +c, +s, +r, +w

  -c, +s, +r, -w

  +c, -s, +r, +w

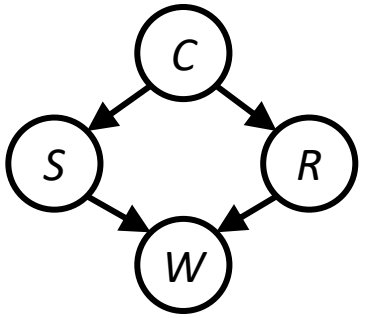
  -c, -s, -r, +w



- What is P(W)?
  - We have counts <+w:4, -w:1>
  - Normalize to get P(W) = <+w:0.8, -w:0.2>
  - 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?)

Rejection Sampling

- Let's say we want P(C)
  - No point in keeping all the samples around
  - Just keep counts of C
- Let's say we want P(C| +s)
  - Keep counts of C when there is also S = +s
  - Ignore (i.e. reject) other samples
  - This is called rejection sampling
- Rejection sampling is consistent (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

Rejection Sampling

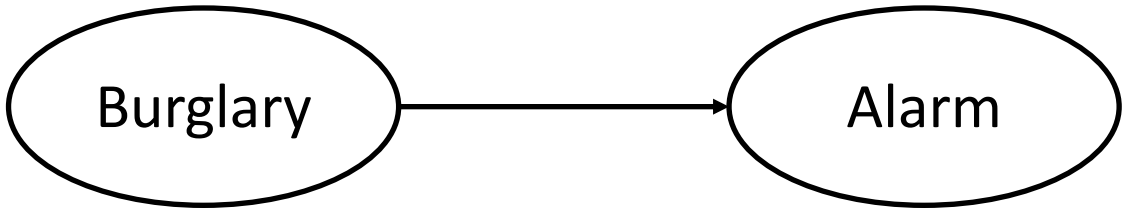
- Given a BN as before and the evidence $E(e_1, \ldots, e_k)$

```
for i = 1 to n do
    x_i = sample(P(X_i | Parents(X_i)))
    if inconsistent(E, x_i) then
        return Null
return (x_1, ..., x_n)
```

- $x_i \notin E \Rightarrow inconsistent(E, x_i)$
- No sample-set is generated when a single sample is rejected

Likelihood Weighting

- What could be problem with rejection sampling?
- If evidence is unlikely, a lot of samples are rejected!
- E.g. P(B|+a)?

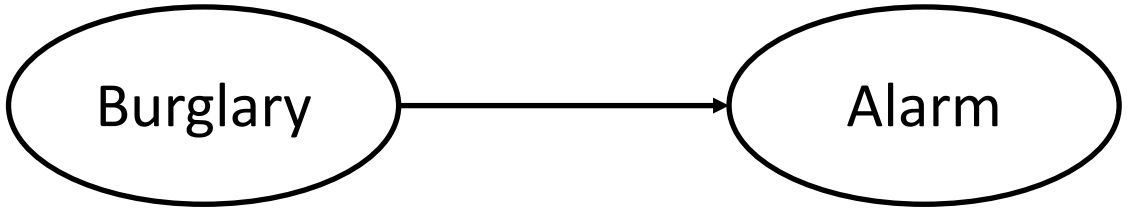


P(+b) = 0.01
P(-b) = 0.99
P(+a|+b) = 0.8
P(-a|+b) = 0.2
P(+a|-b) = 0.05
P(-a|-b) = 0.95

~~-b, -a~~
~~-b, -a~~
-b, +a
~~-b, -a~~
+b, +a
~~-b, -a~~
~~-b, -a~~
~~-b, -a~~
-b, +a
~~-b, -a~~
~~-b, -a~~

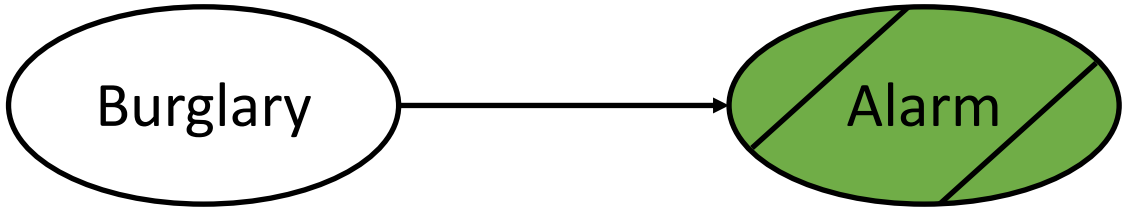
- Evidence does not guide sampling!

Likelihood Weighting



P(+b) = 0.01
P(-b) = 0.99
P(+a|+b) = 0.8
P(-a|+b) = 0.2
P(+a|-b) = 0.05
P(-a|-b) = 0.95

- Idea 1: Fix Evidence variables and sample the rest



-b, +a
-b, +a
-b, +a
-b, +a
+b, +a
-b, +a
-b, +a
-b, +a
-b, +a
-b, +a
-b, +a

- Problem?
- Sample distribution is not consistent!
- Solution: Weight by **probability of evidence** given parents

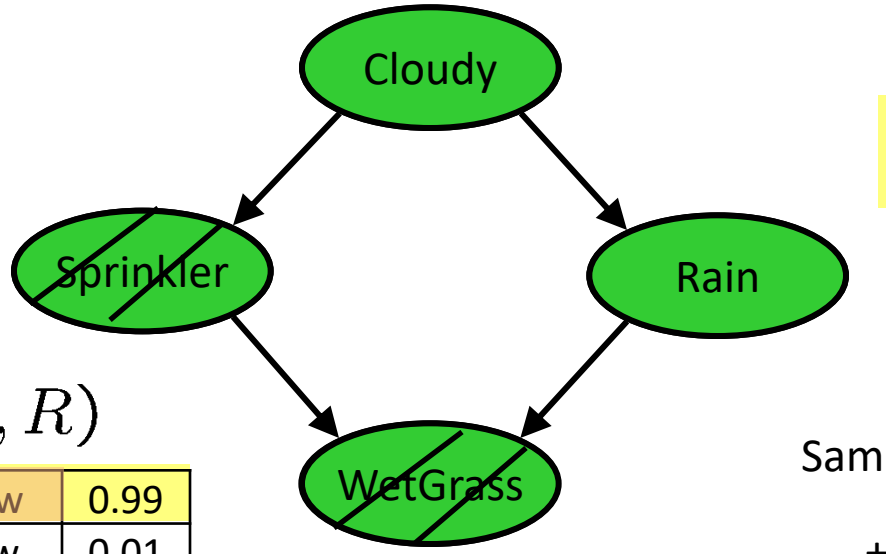
Likelihood Weighting

$P(C)$

+c	0.5
-c	0.5

$P(S|C)$

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



$P(R|C)$

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

$P(W|S,R)$

+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

Samples:

+c, +s, +r, +w

…

$$w = 1.0 \times 0.1 \times 0.99$$

Likelihood Weighting

- Given a BN and the evidence as before

  w = 1.0
  for i = 1 to n do
    if X_i is an evidence variable then
      $X_i = x_i$, where $x_i \in E$ is the observation for X_i
      $w = w \times P(X_i | Parents(X_i))$
    else
      $x_i = sample(P(X_i | Parents(X_i))$
  return $(x_1, \ldots, x_n)$, w

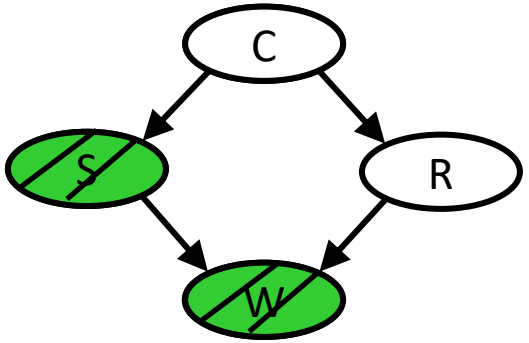
- Instead of keeping counts of the samples, keep the sum of weights (there will be an example a few slides later).
- See also: Figure 14.15 in the book

Probability:
sum of weights for the desired part divided by the total sum

Likelihood Weighting

- Sampling distribution if z sampled and e fixed evidence

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



- Now, samples have weights

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

- Together, weighted sampling distribution is consistent

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

$$= P(\mathbf{z}, \mathbf{e})$$

So far

- Exact inference, even with variable elimination, can be computationally slow

- Approximates inference trades-off accuracy with computational concerns
  - More time higher accuracy

- Basic Idea: Sample from the distribution and use the samples to approximate the desired distribution

So Far

- Prior Sampling: P(Q)
  - Sample away!
  - Used when we want to sample without evidence
- Rejection Sampling: P(Q | e)
  - Reject samples that do not agree with the evidence
  - Inefficient, may reject a lot of samples
- Likelihood Weighting: P(Q | e)
  - Sample while keeping evidence fixed
  - But calculate *weights* such that it becomes consistent

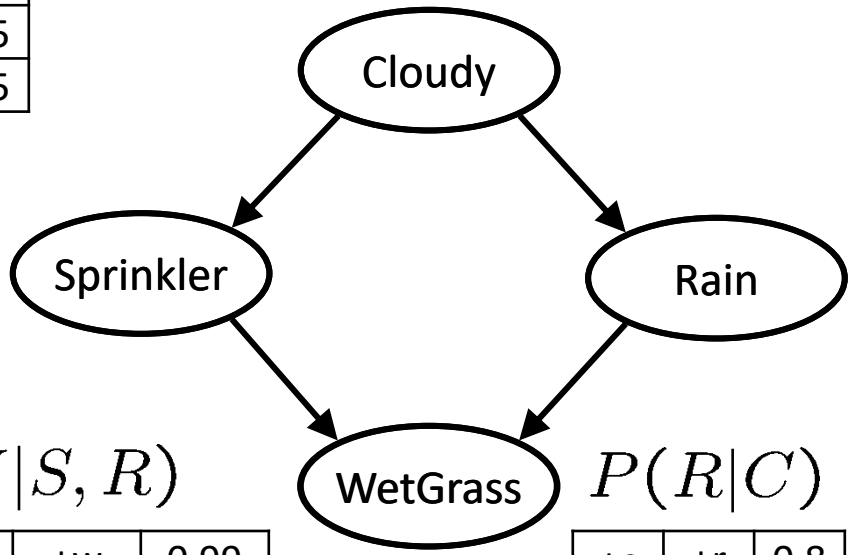
Examples – Prior and Rejection Sampling

$P(C)$

+c	0.5
-c	0.5

$P(S|C)$

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



$P(W|S,R)$

+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

$P(R|C)$

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

Samples (20):

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

Prior Sampling

- $P(+c,+w)$ = ? (0.3735 exact)

$$\frac{N(+c,+w)}{N_{total}} = \frac{6}{20} = 0.3$$

Rejection sampling

- For $(+c,+w)$, which samples are rejected?
- $P(R|+c,+w) = ?$

$(P(+r\,|+c,+w) = 0.9758$ exact)

$$\frac{N(+r,+c,+w)}{N(+c,+w)} = \frac{6}{6} = 1$$

$$\frac{N(-r,+c,+w)}{N(+c,+w)} = \frac{0}{6} = 0$$

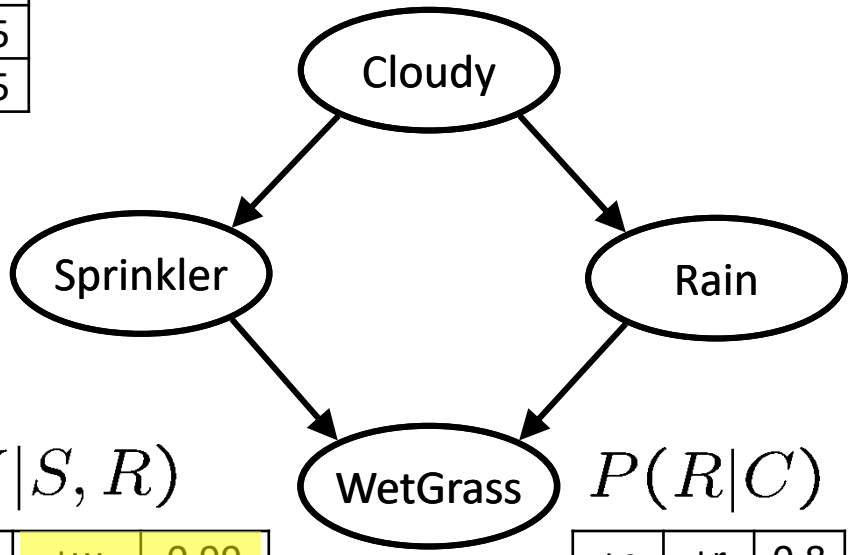
Example – Likelihood Weighing

$P(C)$

+c	0.5
-c	0.5

$P(S|C)$

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



$P(W|S,R)$

+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

$P(R|C)$

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

For evidence $(+c, +w)$ (20):

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

+c,+s,+w,+r: $0.5 \cdot 0.99$, 4
+c,-s,+w,+r: $0.5 \cdot 0.90$, 15
+c,-s,+w,-r: $0.5 \cdot 0.01$, 1
+c,+s,+w,-r: $0.5 \cdot 0.01$, 0

$P(+r \mid +c, +w)$ = ? (0.9758 exact)

$$\frac{4 \cdot 0.5 \cdot 0.99 + 15 \cdot 0.5 \cdot 0.9}{4 \cdot 0.5 \cdot 0.99 + 15 \cdot 0.5 \cdot 0.9 + 1 \cdot 0.5 \cdot 0.01}$$

$$\approx 0.999$$

$P(+s \mid +c, +w)$ = ? (0.1304 exact)

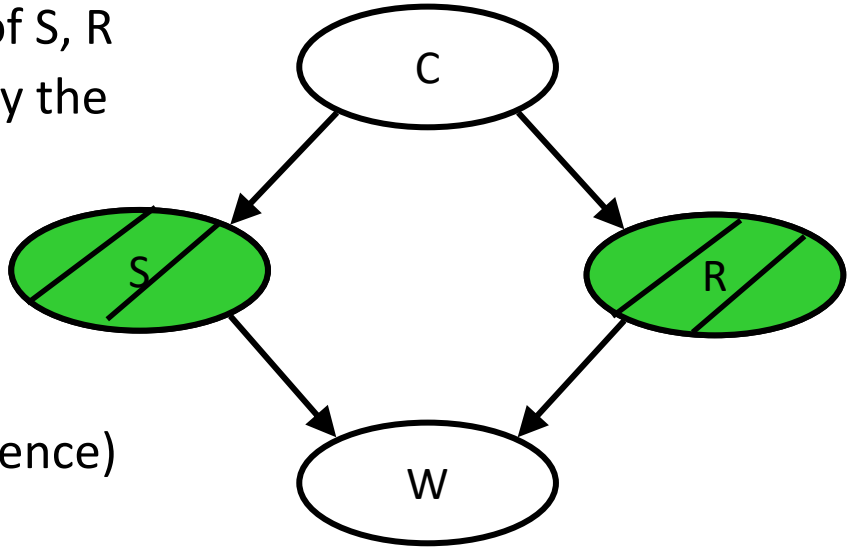
$$\frac{4 \cdot 0.5 \cdot 0.99}{4 \cdot 0.5 \cdot 0.99 + 15 \cdot 0.5 \cdot 0.9 + 1 \cdot 0.5 \cdot 0.01}$$

$$\approx 0.227$$

Better than rejection sampling but 20 samples is not enough for this problem

Likelihood Weighting

- Likelihood weighting is good
  - We have taken evidence into account as we generate the sample
  - E.g. here, W's value will get picked based on the evidence values of S, R
  - More of our samples will reflect the state of the world suggested by the evidence



- Likelihood weighting doesn't solve all our problems
  - Evidence influences the choice of downstream variables, but not upstream ones (C isn't more likely to get a value matching the evidence)

- We would like to consider evidence when we sample every variable

  → Gibbs sampling

Gibbs Sampling

- *Procedure:* Keep track of a full instantiation $x_1, x_2, \ldots, x_n$. 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.

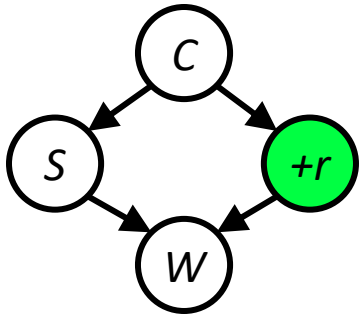
- *Property:* In the limit of repeating this infinitely many times, the resulting sample is coming from the correct distribution

- *Rationale*: Both upstream and downstream variables condition on evidence!

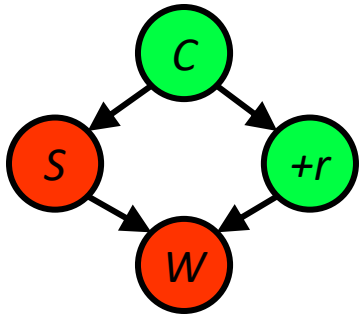
- In contrast, likelihood weighting only conditions on upstream evidence, and hence weights obtained in likelihood weighting can sometimes be very small. Sum of weights over all samples is indicative of how many "effective" samples were obtained, so we want large weights.

Gibbs Sampling Example: P(S | +r)

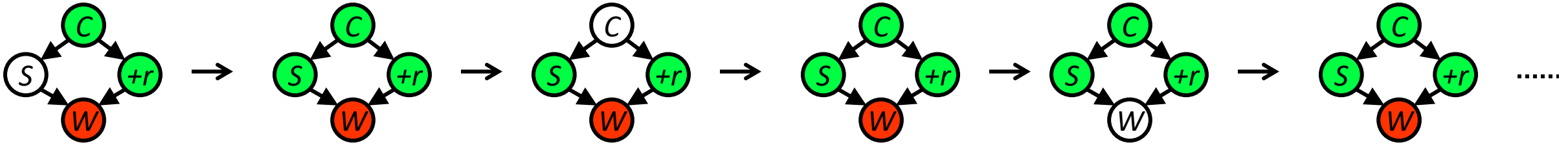
- Step 1: Fix evidence
  - R = +r



- Step 2: Initialize other variables
  - Randomly



- Steps 3: Repeat
  - Choose a non-evidence variable X
  - Resample X from P(X | all other variables)

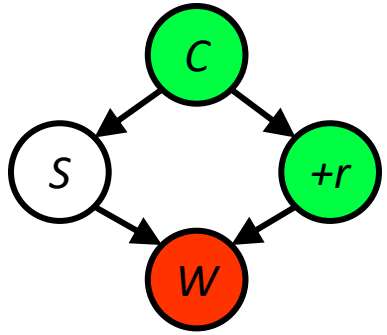


Sample from $P(S|+c,-w,+r)$ Sample from $P(C|+s,-w,+r)$ Sample from $P(W|+s,+c,+r)$

Efficient Resampling of One Variable

- Sample from P(S | +c, +r, -w)

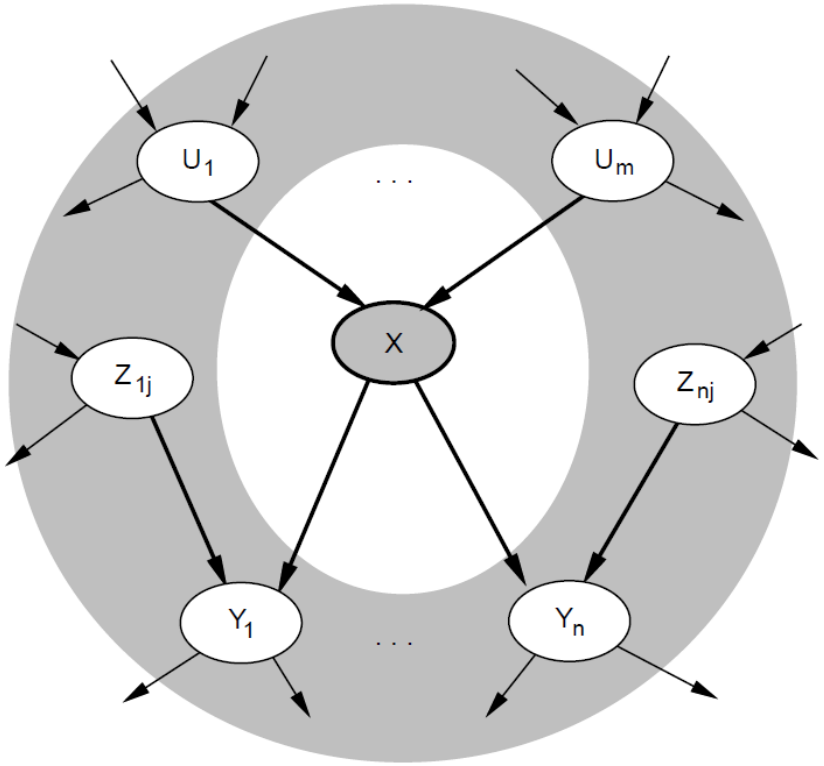
$$
\begin{aligned}
P(S|+c,+r,-w) &= \frac{P(S,+c,+r,-w)}{P(+c,+r,-w)} \\
&= \frac{P(S,+c,+r,-w)}{\sum_s P(s,+c,+r,-w)} \\
&= \frac{P(+c)P(S|+c)P(+r|+c)P(-w|S,+r)}{\sum_s P(+c)P(s|+c)P(+r|+c)P(-w|s,+r)} \\
&= \frac{P(+c)P(S|+c)P(+r|+c)P(-w|S,+r)}{P(+c)P(+r|+c)\sum_s P(s|+c)P(-w|s,+r)} \\
&= \frac{P(S|+c)P(-w|S,+r)}{\sum_s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



- Many things cancel out – only CPTs with S remain!
- More generally: only CPTs that have resampled variable need to be considered, and joined together – Markov Blanket!

Recall: Markov Blanket

- Each node is conditionally independent of all others given its *Markov Blanket*
- Markov Blanket: parents + children + children's parents



Gibbs Sampling

- Given a BN, the evidence as before and a full instantiation $(x_1, \ldots, x_n)$

  `Pick a random non-evidence variable,` X_i

  $x_i = sample(P(X_i|MarkovBlanket(X_i)))$

  `return` $(x_1, \ldots, x_n)$

- Then calculate the probabilities using counts as in rejection sampling (no weights!)
- See also: Figure 14.16 in the book

Further Reading on Gibbs Sampling*

- Gibbs sampling produces sample from the query distribution P(Q | e) in the limit of re-sampling infinitely often
- Since we do not have infinite time, we obviously stop at some point. Furthermore, a "burn in" period for Gibbs sampling is required where an initial batch of the samples are thrown out.
- Gibbs sampling is a special case of more general methods called Markov chain Monte Carlo (MCMC) methods
  - Metropolis-Hastings is one of the more famous MCMC methods (in fact, Gibbs sampling is a special case of Metropolis-Hastings)
  - For a more formal treatment, we need to learn about Markov Models
- You may read/hear about Monte Carlo methods – they're just sampling

Applications of MCMC Methods

- Inference in BNs, Bayesian methods in general
- Computer Vision (e.g. texture distribution)
- Robotics (e.g. localization)
- Machine learning in general (PGMs, Bayesian Models)
- Simulating complex stochastic systems, e.g. weather prediction
- Statistical physics, approximate counting, computing volumes and integrals, and combinatorial optimization

Summary

- Prior Sampling: P(Q)
  - Sample away!
  - Used when we want to sample without evidence
- Rejection Sampling: P(Q | e)
  - Reject samples that do not agree with the evidence
  - Inefficient, may reject a lot of samples
- Likelihood Weighting: P(Q | e)
  - Sample while keeping evidence fixed
  - But calculate *weights* such that it becomes consistent
  - Does not influence “upstream” variables
- Gibbs Sampling: P(Q | e)
  - Keep evidence variables fixed
  - Chose a non-evidence variable and resample