343H: Honors AI

Lecture 17: Bayes Nets Sampling

3/25/2014

Kristen Grauman

UT Austin

Slides courtesy of Dan Klein, UC Berkeley

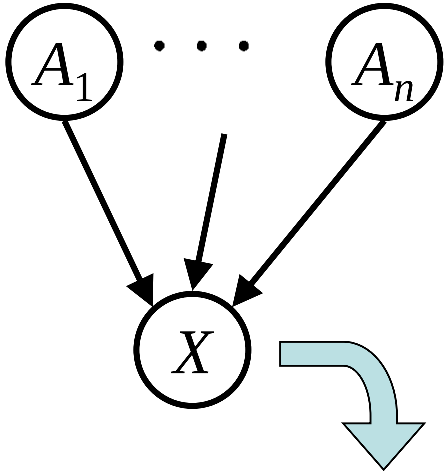
Road map: Bayes' Nets

- ✓ Representation
- ✓ Conditional independences
- Probabilistic inference
  - ✓ Enumeration (exact, exponential complexity)
  - ✓ Variable elimination (exact, worst-case exponential complexity, often better)
  - ✓ Inference is NP-complete
  - Sampling (approximate)
- Learning Bayes' Nets from data

Recall: Bayes' Net Representation

- A directed, acyclic graph, one node per random variable
- A conditional probability table (CPT) for each node
  - A collection of distributions over X, one for each combination of parents' values

$$P(X|a_1 \ldots a_n)$$



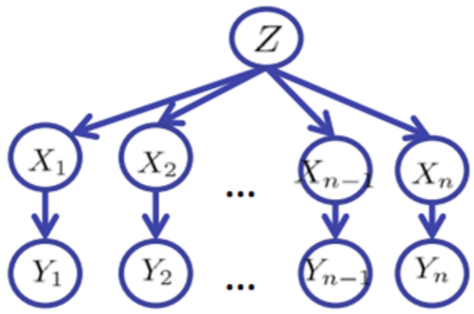
$$P(X|A_1 \ldots A_n)$$

- Bayes' nets implicitly encode joint distributions
  - As a product of local conditional distributions

$$P(x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} P(x_i|parents(X_i))$$

Last time: Variable elimination

- Interleave joining and marginalizing
- d^k entries computed for a factor with k variables with domain sizes d
- Ordering of elimination of hidden variables can affect size of factors generated
- Worst case: running time exponential in the size of the Bayes' net.



Sampling

- Sampling is a lot like repeated simulation
  - Predicting the weather, basketball games,…
- Basic idea:
  - Draw N samples from a sampling distribution S
  - Compute an approximate posterior probability
  - Show this converges to the true probability P
- Why sample?
  - Inference: getting a sample is faster than computing the right answer (e.g. with variable elimination)
  - Learning: get samples from a distribution you don't know

Sampling

- Sampling from a 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

C	P(C)
red	0.6
green	0.1
blue	0.3

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

If random() returns u=0.83, then our sample C = blue.

Sampling in Bayes' Nets

- Prior sampling
- Rejection sampling
- Likelihood weighting
- Gibbs sampling

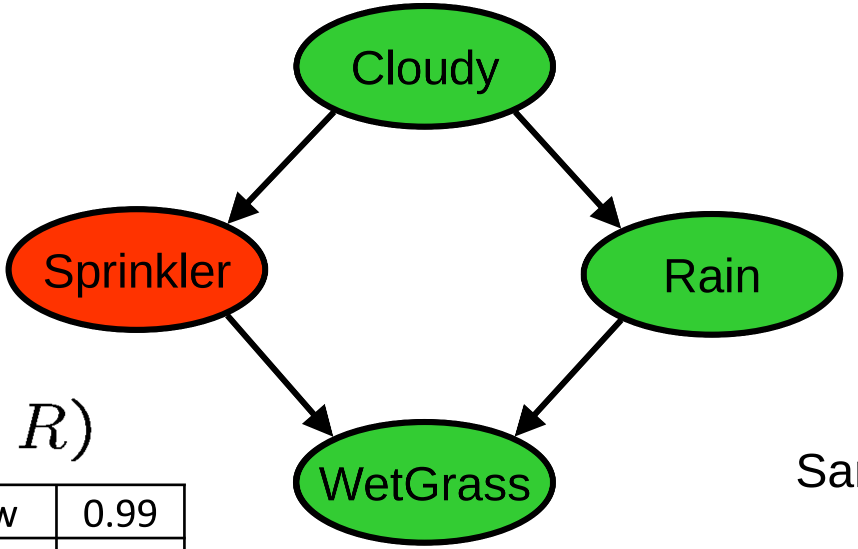
Prior Sampling

$P(C)$

+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

- For i=1, 2, …, n
  - Sample x_i from $P(X_i \mid Parents(X_i))$
- Return $(x_1, x_2, \ldots, x_n)$

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 $$\begin{aligned} \lim_{N\to\infty} \widehat{P}(x_1, \ldots, x_n) &= \lim_{N\to\infty} N_{PS}(x_1, \ldots, x_n)/N \\ &= S_{PS}(x_1, \ldots, x_n) \\ &= P(x_1 \ldots x_n) \end{aligned}$$

- I.e., the sampling procedure is consistent

Example

- First: Get a bunch of samples from the BN:

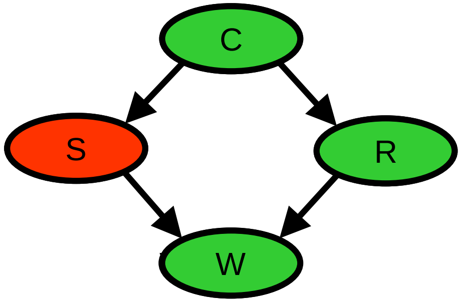
  +c, -s, +r, +w

  +c, +s, +r, +w

  -c, +s, +r, -w

  +c, -s, +r, +w

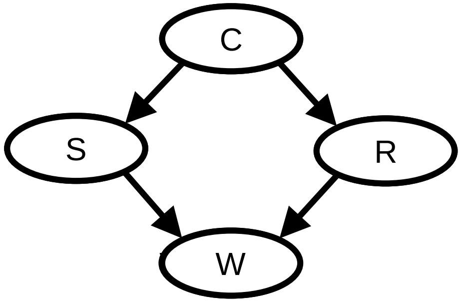
  -c, -s, -r, +w



- Example: we want to know P(W)
  - We have counts <+w:4, -w:1>
  - Normalize to get approximate 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 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

Rejection sampling

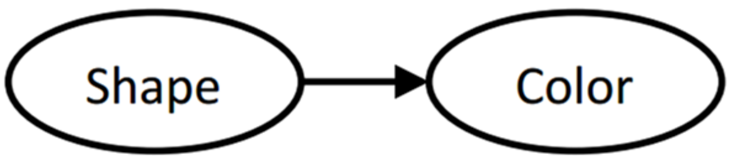
- IN: evidence instantiation
- For i=1, 2, …, n
  - Sample x_i from $P(X_i \mid Parents(X_i))$
  - If x_i not consistent with evidence
    - Reject: Return, and no sample is generated in this cycle
- Return $(x_1, x_2, \ldots, x_n)$

Sampling Example

- There are 2 cups.
  - The first contains 1 penny and 1 quarter
  - The second contains 2 quarters
- Say I pick a cup uniformly at random, then pick a coin randomly from that cup. It's a quarter (yes!).
- What is the probability that the other coin in that cup is also a quarter?

Likelihood weighting

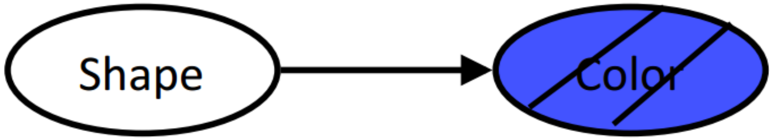
- Problem with rejection sampling:
  - If evidence is unlikely, you reject a lot of samples
  - Evidence not exploited as you sample
  - Consider P(Shape | blue)



~~pyramid, green~~
~~pyramid, red~~
sphere, blue
~~cube, red~~
~~sphere, green~~

Likelihood weighting

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



pyramid, blue
pyramid, blue
sphere, blue
cube, blue
sphere, blue

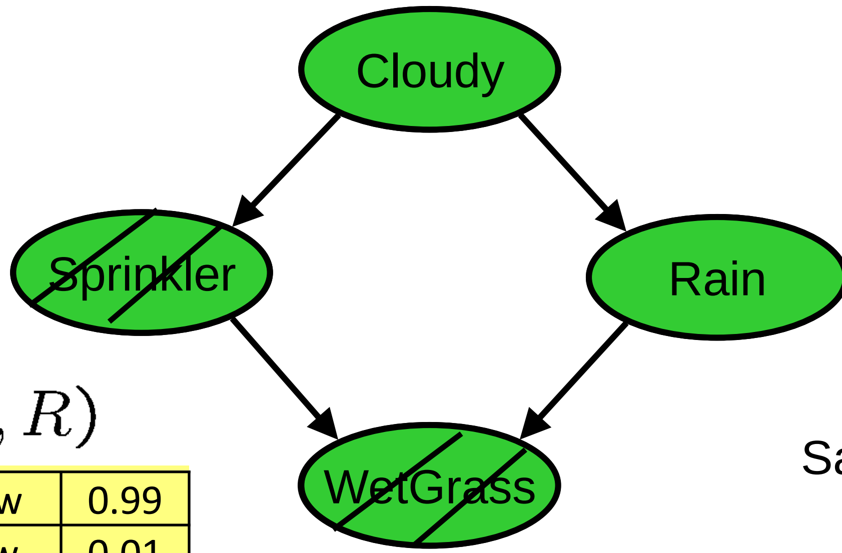
Likelihood Weighting

$P(C)$

+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

…

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

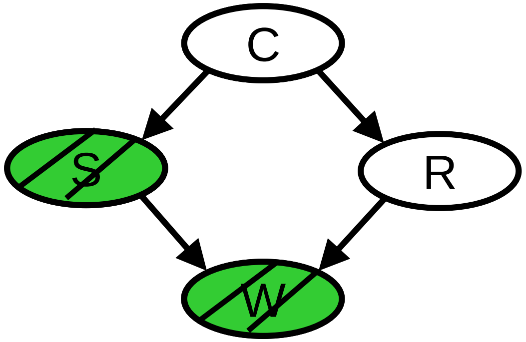
Likelihood weighting

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

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

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

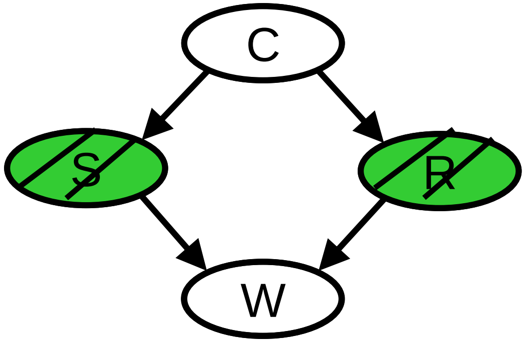
- Together, weighted sampling distribution is consistent

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

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

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

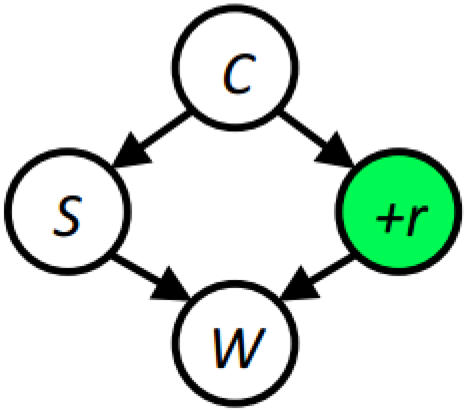
- **Procedure**:
  - Keep track of a full instantiation $x_1, x_2, \ldots x_n$.
  - Start with an arbitrary instantiation consistent with the evidence.
  - Sample one variable at a time, conditioned on all the rest, but keep evidence fixed.
  - Keep repeating this for a long time.
- **Property**:
  - In the limit of repeating this infinitely many times, the resulting sample is coming from the correct distribution.

Gibbs sampling

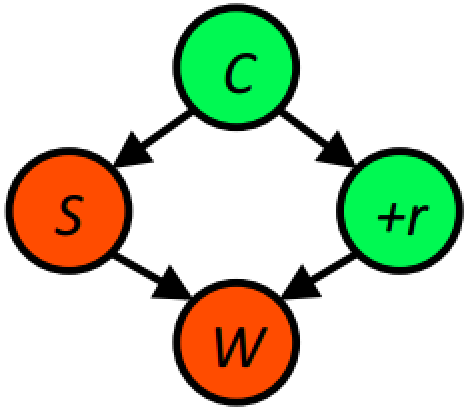
- **Rationale**:
  - Both upstream and downstream variables condition on the evidence.
- **In contrast**:
  - Likelihood weighting only conditions on upstream evidence, hence weights obtained in likelihood weighting can sometimes be very small.
  - Sum of weights over all samples is indicative of how many “effective” samples were obtained, so we want high weight.

Gibbs sampling example: P(S | +r)

- Step 1: Fix evidence
  - R = +r

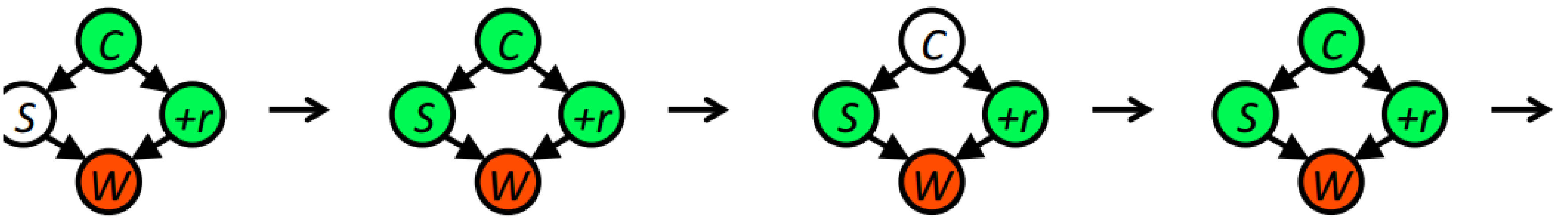


- Step 2: Initialize other variables
  - Randomly

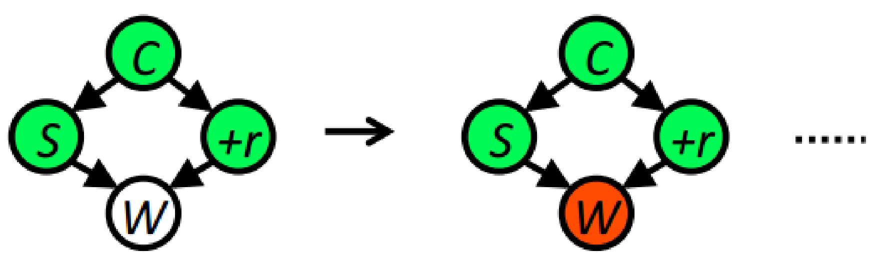


Gibbs sampling example: P(S | +r)

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



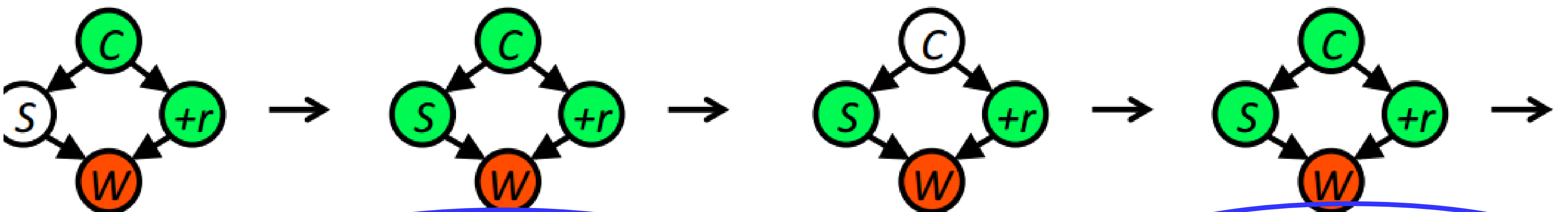
Sample from $P(S|+c,-w,+r)$ Sample from $P(C|+s,-w,+r)$



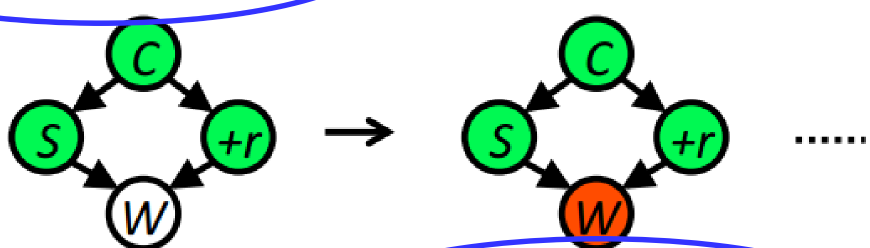
Sample from $P(W|+s,+c,+r)$

Gibbs sampling example: P(S | +r)

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



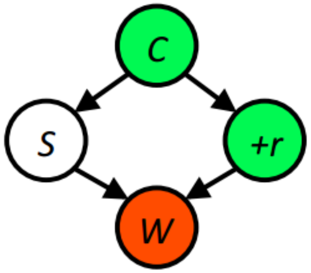
Sample from $P(S|+c,-w,+r)$ Sample from $P(C|+s,-w,+r)$



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, joined together.

Gibbs and MCMC

- Gibbs sampling produces sample from query distribution P(Q | e) in limit of resampling infinitely often
- Gibbs is a special case of more general methods called Markov chain Monte Carlo (MCMC) methods

Bayes’ Net sampling summary

- Prior sampling P
- Rejection sampling P(Q | e)
- Likelihood weighting P(Q | e)
- Gibbs sampling P(Q | e)

Reminder

- Check course page for
  - Contest (today)
  - PS4 (Thursday)
  - Next week’s reading