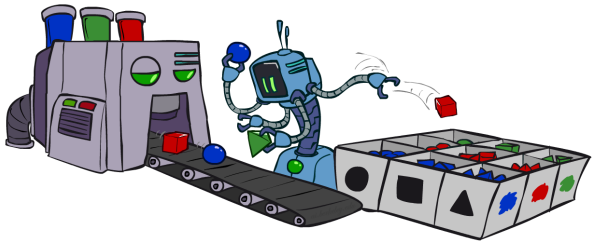
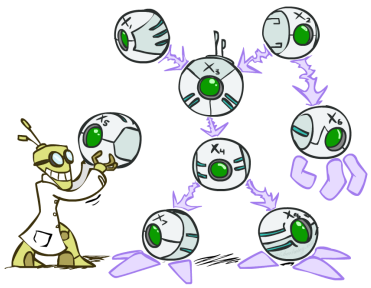
CS 188: Artificial Intelligence



Bayes' Nets: Sampling

Bayes' Net Representation



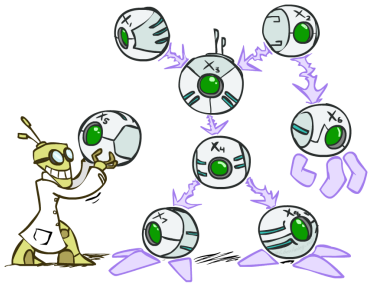
$$P(X|a_1 \dots a_n)$$

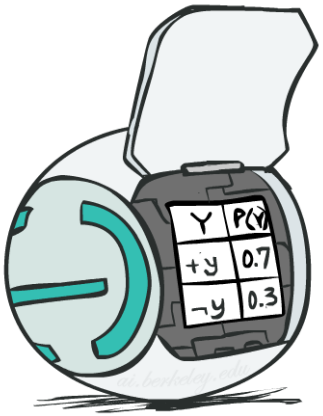
A directed, acyclic graph, one node per random variable

A conditional probability table (CPT) for each node

- A collection of distributions over X, one for each combination of parents' values

Bayes' Net Representation





Bayes' nets implicitly encode joint distributions

- As a product of local conditional distributions
- Probability of a full assignment in BN is product of relevant conditionals:

$$P(x_1, \ldots, x_n) = \Pi_{i=1}^{n} P(x_i | Parents(X_i))$$

- Less work than chain rule (valid for all distributions):

$$P(x_1, \ldots, x_n) = \Pi_{i=1}^{n} P(x_i | x_1, \ldots, x_i)$$

Bayes' Nets

Representation ✓

D-separation. ✓

Probabilistic Inference. ✓

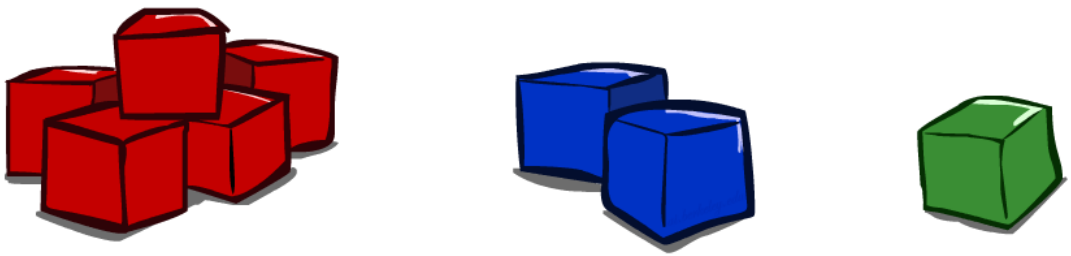
- Enumeration (exact, exponential complexity). ✓
- Variable elimination

(exact, worst-case exponential complexity, often better) . ✓

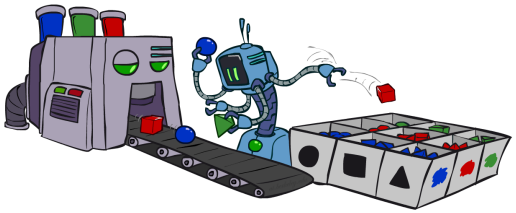
- Inference is NP-complete. ✓
- Sampling (approximate) (Next up.)

Learning Bayes' Nets from Data. (Later.)

Approximate Inference: Sampling



Sampling



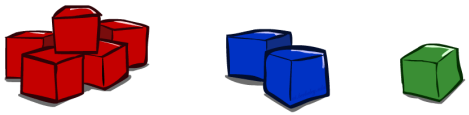
Basic idea.

- 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)

Sampling Basics



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

Sampling from given distribution

- Step 1: Sample u uniformly from $[0, 1)$
- E.g. random() in python
- Step 2: Convert sample u into outcome ω using sub-interval of $[0, 1)$ of size $P(\omega)$.

Example:

C	P(C)
red	0.6
green	0.1
blue	0.3

- If $u = 0.83$, our sample is $C = blue$
- E.g, after sampling 8 times:
  5 red$, 1$ green$, 2$ blue

Sampling in Bayes' Nets

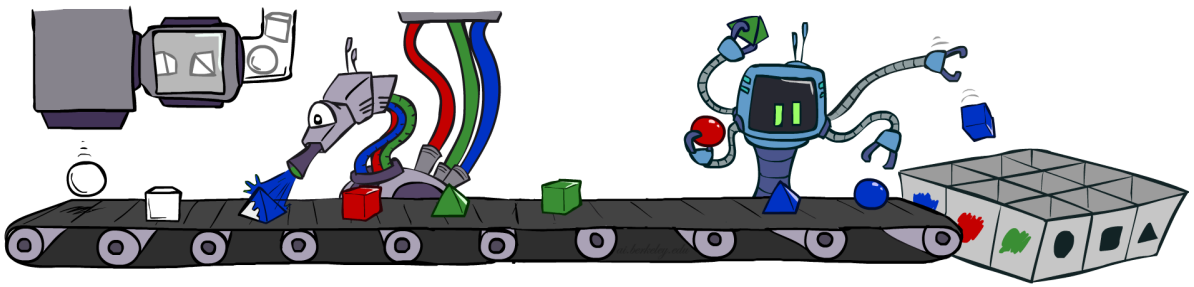
Prior Sampling

Rejection Sampling

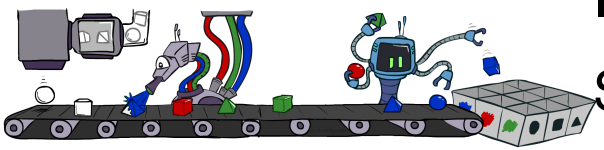
Likelihood Weighting

Gibbs Sampling

Prior Sampling



Prior Sampling



Ignore evidence.

Sample from the joint probability.

Do inference by counting the right samples.

Prior Sampling

Random Variables: C,S,R,W

$P(S|C)$

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

$P(C)$

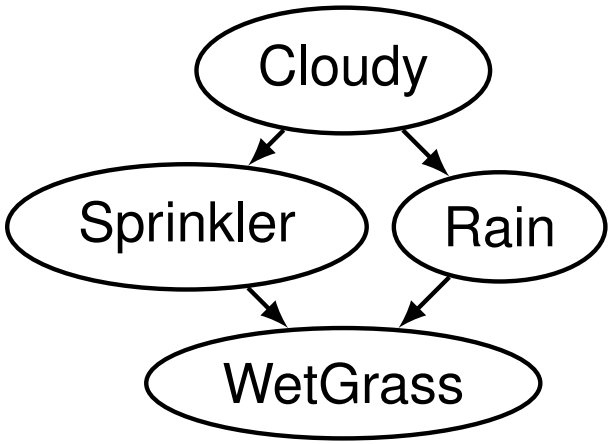
+c	0.5
-c	0.5

$P(R|C)$

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

$P(W|S,R)$

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



Samples:

+c, -s, +r, +w

-c, +s, -r, +w

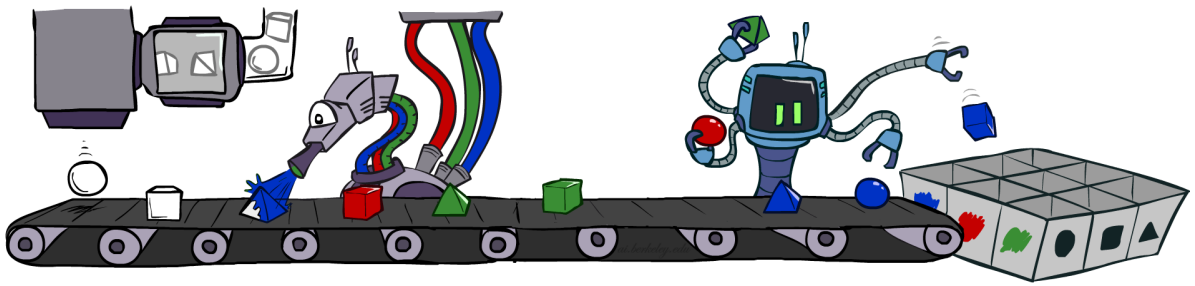
...

Prior Sampling

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

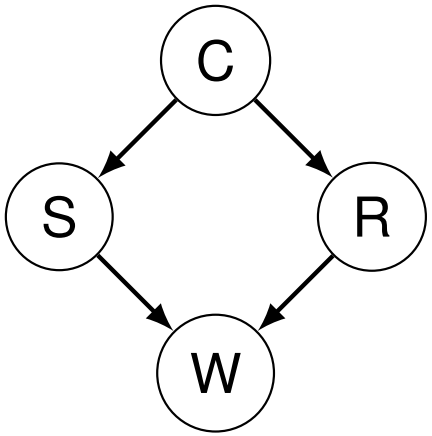
Sample x_i from $P(X_i | Parents(X_i))$

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



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

If we want to know $P(W)$

- 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?)

Prior Sampling Analysis

This process generates samples with probability:

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

i.e. the BN's joint probability

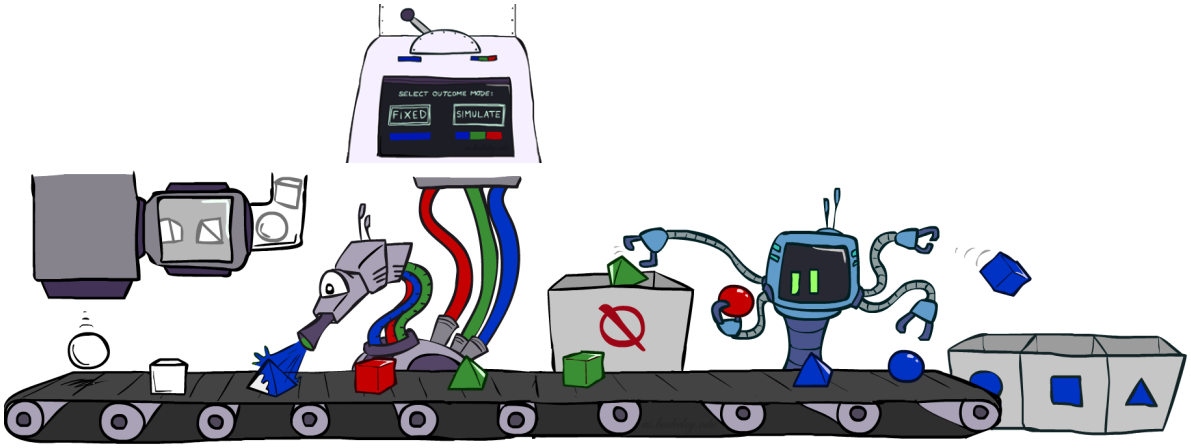
Let the number of samples of an event be $N_{PS}(x_1, \ldots, x_n)$

Then

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

I.e., the sampling procedure is consistent

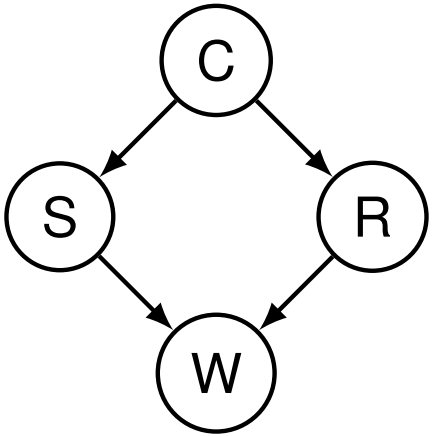
Rejection Sampling



Rejection Sampling

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

- Tally C outcomes, 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

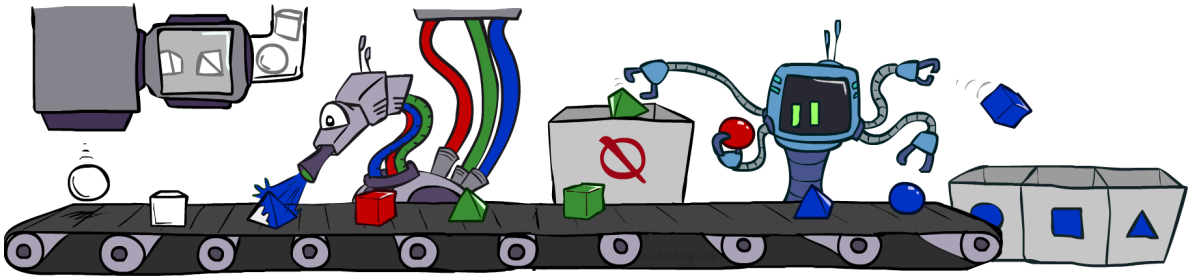
Rejection Sampling

IN: evidence instantiation

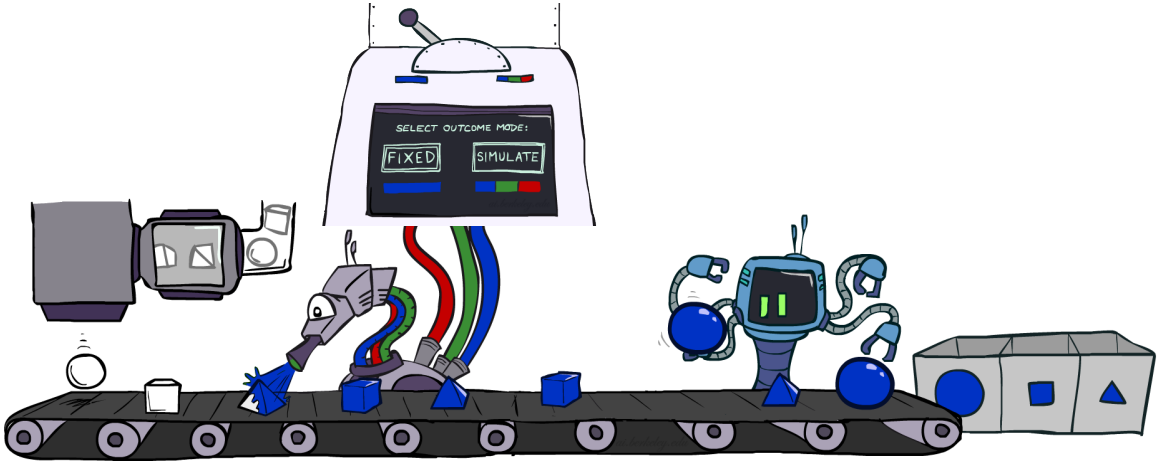
For $i = 1, 2, \ldots, n$

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

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



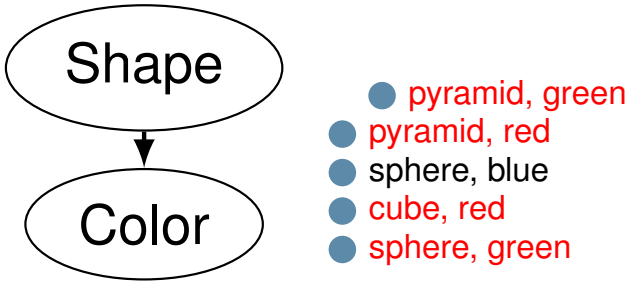
Likelihood Weighting

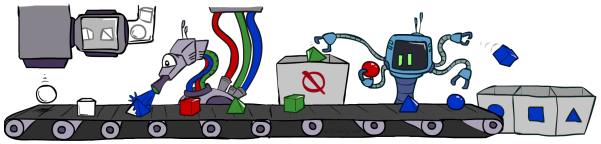


Likelihood Weighting

Problem with rejection sampling:

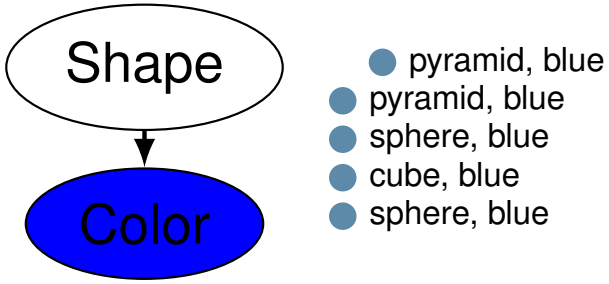
- If evidence unlikely, rejects lots of samples
- Evidence not exploited as you sample
- Consider $P(Shape|blue)$

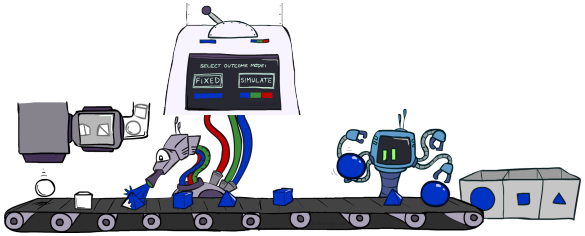




Idea: fix evidence variables and sample the rest

- Problem: sample distribution not consistent!
- Solution: weight by probability of evidence given parents





Likelihood weighting

Random Variables: C,S,R,W.

Observed: S=+s, W=+w

$P(S|C)$

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

$P(C)$

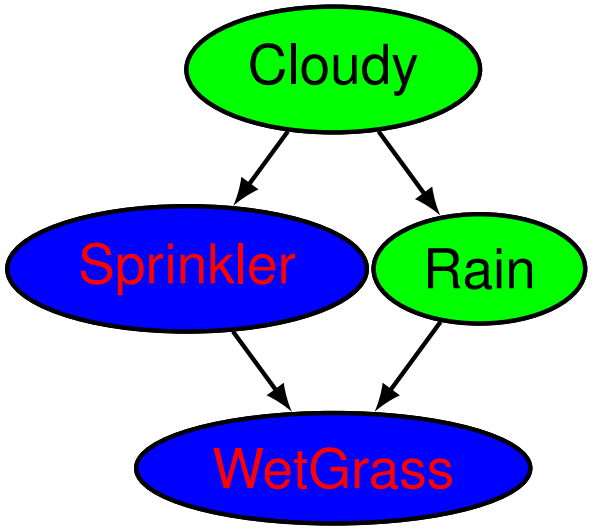
+c	0.5
-c	0.5

$P(R|C)$

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

$P(W|S,R)$

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



Sample:

+c, +s, +r, +w

Weight:

$w = 1.0 \times .1 \times .99$

.1 from +s

.99 from +w

Likelihood Weighting

IN: evidence instantiation

$w = 1.0$

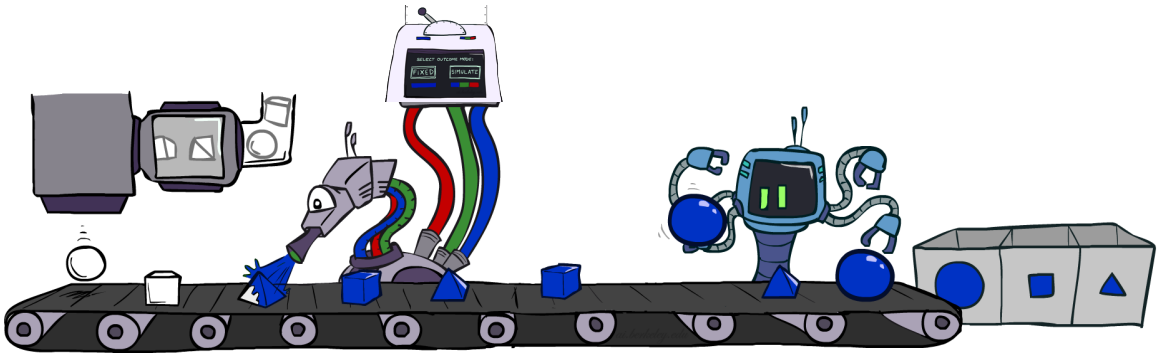
for $i = 1, 2, \ldots, n$ if X_i is an evidence variable

X_i = observation x_i for X_i
Set $w = w \times P(x_i | Parents(X_i))$

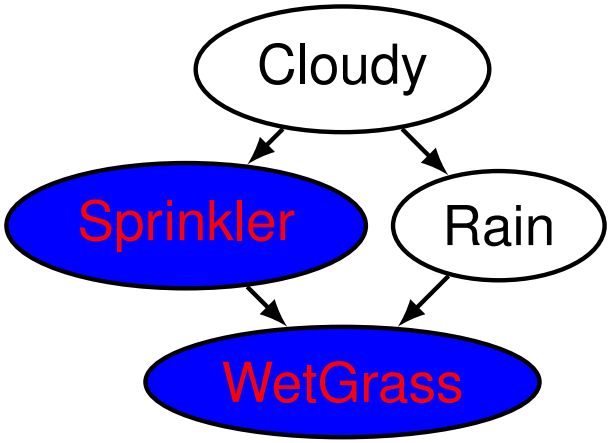
else

Sample x_i from $P(X_i | Parents(X_i))$

return $(x_1, x_2, \ldots, x_n), w$



Likelihood Weighting



Sampling distribution if z sampled and e fixed evidence

$$S_{WS}(z,e) = \Pi_{i=1}^{n} P(z_i | Parents(Z_i))$$

Now, samples have weights:

$$w(z,e) = \Pi_{i=1}^{n} P(e_i | Parents(E_i))$$

Together, weighted sampling distribution is consistent:

$$S_{WS(z,e)} \cdot w(z,e) = \Pi_{i=1}^{n} P(z_i | Parents(Z_i)) \Pi_{i=1}^{n} P(e_i | Parents(E_i))$$
$$= P(z,e)$$

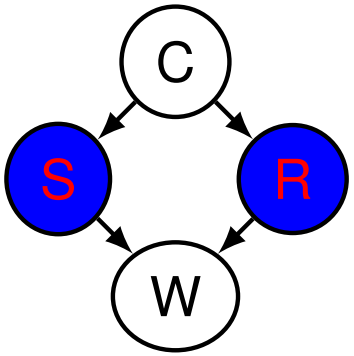
Estimating Probabilities

If we want to know $P(R|+s,+w)$:

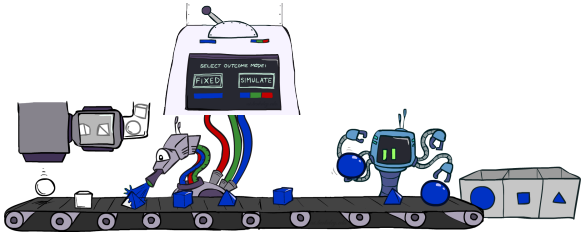
- Same as before, we have samples for +r and –r.
- Instead of just counting, take their weights into account!

Likelihood Weighting

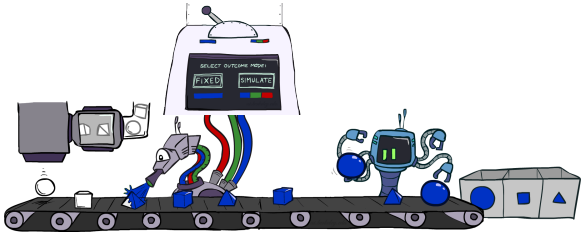
Likelihood weighting is good



- Uses evidence during sample generation.
- E.g. W's value gets picked based on evidence values of S, R
- More samples reflect world state suggested by evidence



Likelihood weighting doesn't solve all our problems

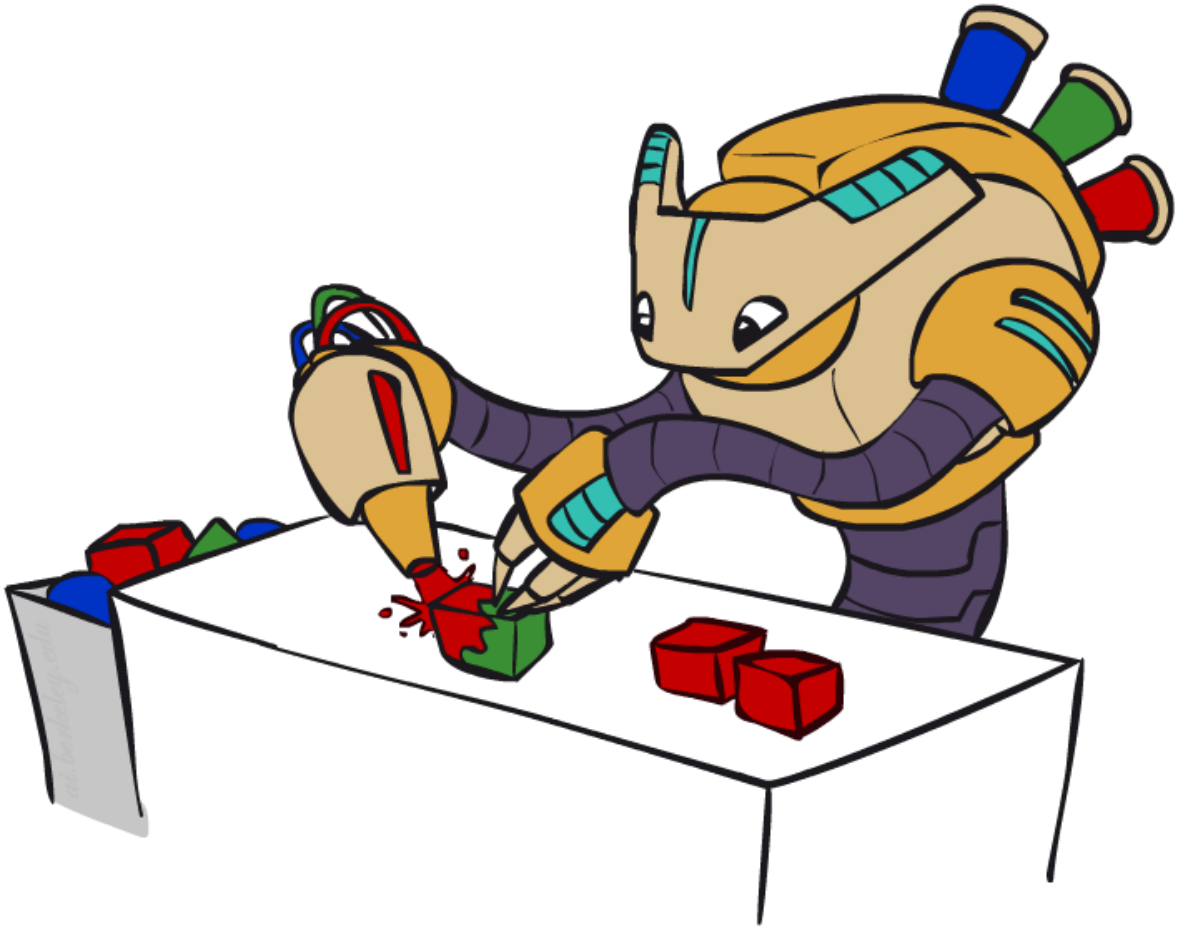


- Evidence influences choice of downstream variables, but not upstream ones (C isn't more likely to get a value matching the evidence)

Like to consider evidence when sample any variable.

Gibbs sampling.

Gibbs Sampling



Gibbs Sampling

Procedure: keep track of a full instantiation $x_1, x_2, \ldots, x_n$.

Start 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.

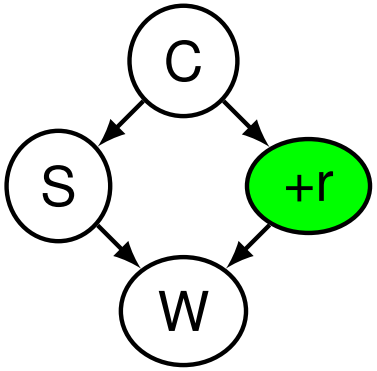
Hence weights obtained in likelihood weighting can sometimes be very small. E.g., make bad choice for C.

Sum of weights over all samples is indicative of how many "effective" samples were obtained, so want high weight.

Gibbs Sampling Example: $P(S|+r)$

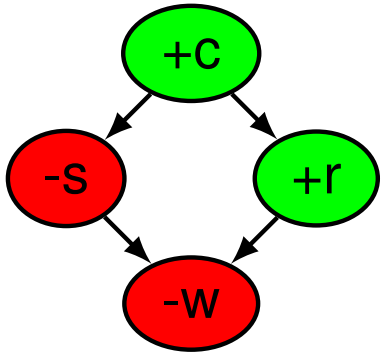
Step 1: Fix evidence

- R = +r



Step 2: Initialize other variables

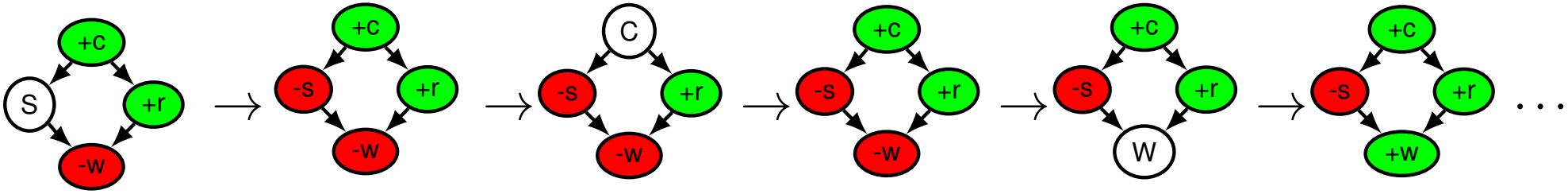
- Randomly



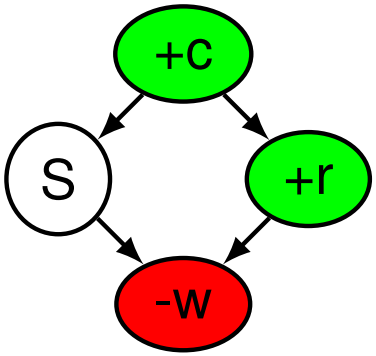
Steps 3: Repeat

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

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



Efficient Resampling of One Variable



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

$$P(S|+c,+r,-w) = \frac{P(S,+c,+r,-w)}{P(+c,+r,-w)}$$

$$= \frac{P(S,+c,+r,-w)}{\sum_s P(s,+c,+r,-w)}$$

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

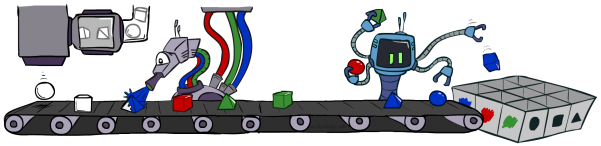
$$= \frac{P(S|+c)P(-w|+r,S)}{\sum_s P(s|+c)P(-w|+r,s)}$$

Many things cancel out – only CPTs with S remain!

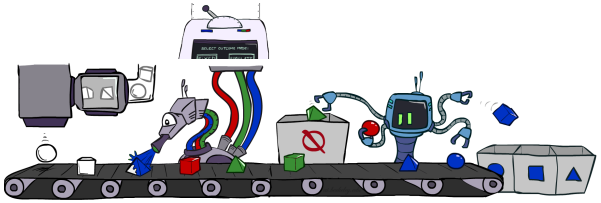
More generally: only CPTs that have resampled variable need to be considered, and joined together

Bayes' Net Sampling Summary

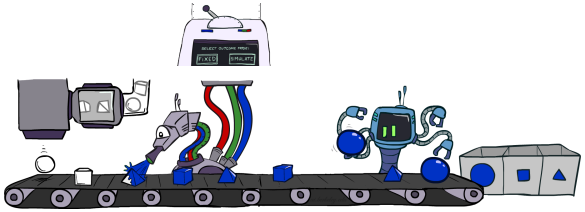
Prior Sampling P



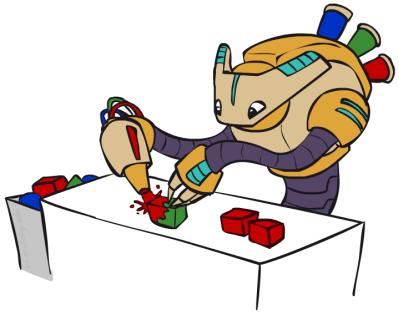
Rejection Sampling $P(Q|e)$



Likelihood Weighting $P(Q|e)$



Gibbs Sampling $P(Q|e)$



Further Reading on Gibbs Sampling*

Gibbs sampling produces sample from the query distribution $P(Q|e)$ in limit of re-sampling infinitely often

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)

Read about Monte Carlo methods – they're sampling!