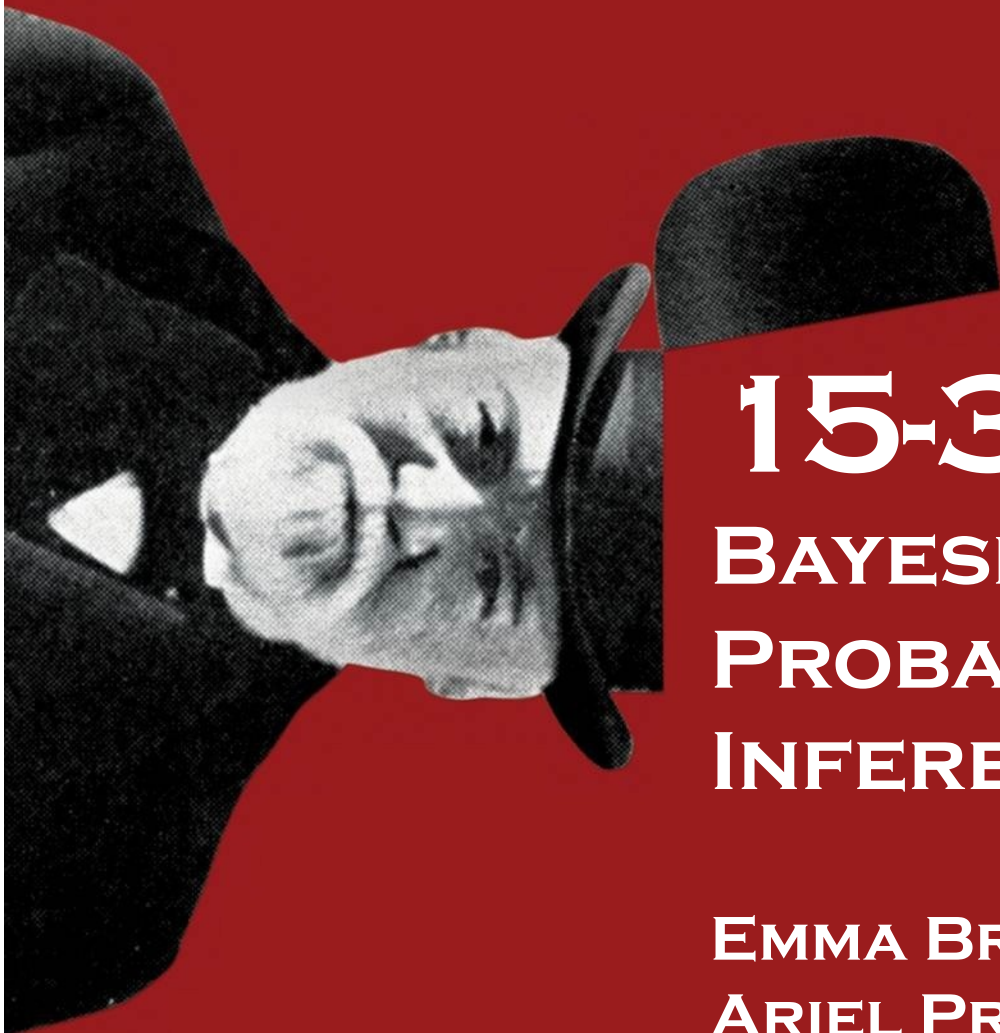


15-381/781

Bayesian Nets & Probabilistic Inference

Emma Brunskill (this time)
Ariel Procaccia

With thanks to Dan Klein (Berkeley), Percy Liang (Stanford) and Past 15-381 Instructors for some slide content, and Russell & Norvig

What You Should Know

- Define probabilistic inference
- How to define a Bayes Net given a real example
- How joint can be used to answer any query
- Complexity of exact inference
- Approximation inference (direct, likelihood, Gibbs)
  - Be able to implement and run algorithm
  - Compare benefits and limitations of each

Bayesian Network

- Compact representation of the joint distribution
- Conditional independence relationships explicit
    - Each var conditionally independent of all its non-descendants in the graph given the value of its parents

Joint Distribution Ex.

- Variables: Cloudy, Sprinkler, Rain, Wet Grass
- Domain of each variable: 2 (true or false)
- Joint encodes probability of all combos of variables & values

P(Cloudy=false & Sprinkler = true & Rain = false & WetGrass = True)

+c	+s	+r	+w	.01
+c	+s	+r	-w	.01
+c	+s	-r	+w	.05
+c	+s	-r	-w	.1
+c	-s	+r	+w	#
+c	-s	+r	-w	#
+c	-s	-r	+w	#
+c	-s	-r	-w	#
-c	+s	+r	+w	#
-c	+s	+r	-w	#
-c	+s	-r	+w	#
-c	+s	-r	-w	#
-c	-s	+r	+w	#
-c	-s	+r	-w	#
-c	-s	-r	+w	#
-c	-s	-r	-w	#

Joint as Product of Conditionals (Chain rule)

+c	+s	+r	+w	.01
+c	+s	+r	-w	.01
+c	+s	-r	+w	.05
+c	+s	-r	-w	.1
+c	-s	+r	+w	#
+c	-s	+r	-w	#
+c	-s	-r	+w	#
+c	-s	-r	-w	#
-c	+s	+r	+w	#
-c	+s	+r	-w	#
-c	+s	-r	+w	#
-c	+s	-r	-w	#
-c	-s	+r	+w	#
-c	-s	+r	-w	#
-c	-s	-r	+w	#
-c	-s	-r	-w	#

=

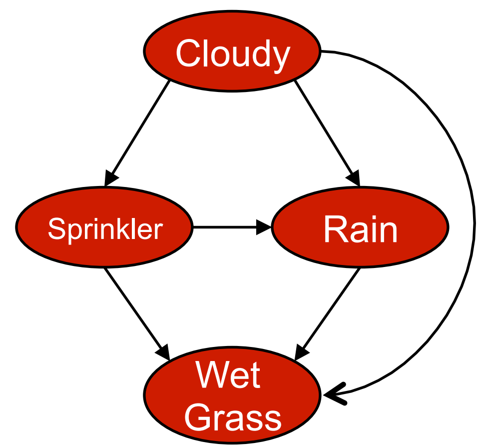
P(WetGrass|Cloudy,Sprinkler,Rain)*
P(Rain|Cloudy,Sprinkler)*
P(Sprinkler|Cloudy)*
P(Cloudy)

Joint as Product of Conditionals

+c	+s	+r	+w	.01
+c	+s	+r	-w	.01
+c	+s	-r	+w	.05
+c	+s	-r	-w	.1
+c	-s	+r	+w	#
+c	-s	+r	-w	#
+c	-s	-r	+w	#
+c	-s	-r	-w	#
-c	+s	+r	+w	#
-c	+s	+r	-w	#
-c	+s	-r	+w	#
-c	+s	-r	-w	#
-c	-s	+r	+w	#
-c	-s	+r	-w	#
-c	-s	-r	+w	#
-c	-s	-r	-w	#

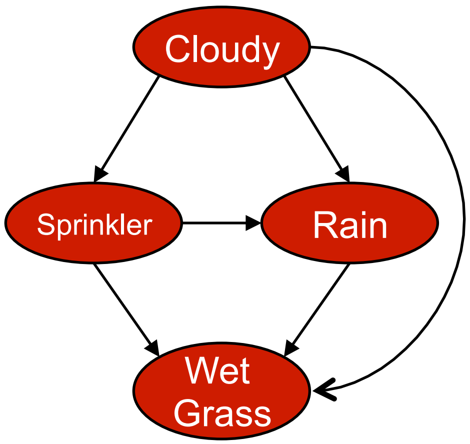
=

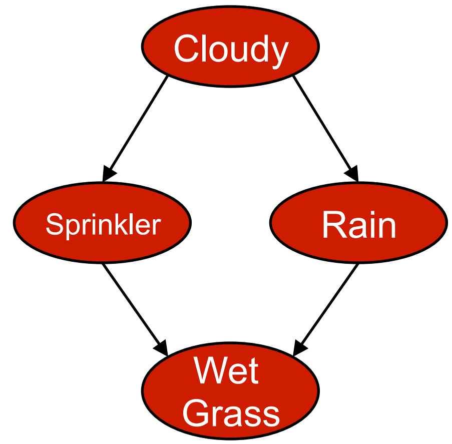
P(WetGrass|Cloudy,Sprinkler,Rain)*
P(Rain|Cloudy,Sprinkler)*
P(Sprinkler|Cloudy)*
P(Cloudy)



…but there may be additional conditional independencies

What if some variables are conditionally indep?





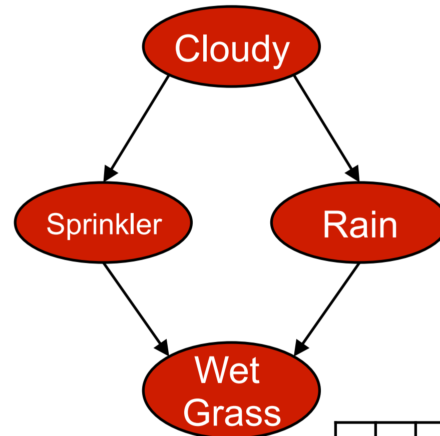
Explicitly shows any conditional independencies

Conditional Independencies

+c	+s	+r	+w	.01
+c	+s	+r	-w	.01
+c	+s	-r	+w	.05
+c	+s	-r	-w	.1
+c	-s	+r	+w	#
+c	-s	+r	-w	#
+c	-s	-r	+w	#
+c	-s	-r	-w	#
-c	+s	+r	+w	#
-c	+s	+r	-w	#
-c	+s	-r	+w	#
-c	+s	-r	-w	#
-c	-s	+r	+w	#
-c	-s	+r	-w	#
-c	-s	-r	+w	#
-c	-s	-r	-w	#

→

+c	0.5
-c	0.5



+c	+s	.1
+c	-s	.9
-c	+s	.5
-c	-s	.5

+c	+r	.8
+c	-r	.2
-c	+r	.2
-c	-r	.8

+s	+r	+w	.99
+s	+r	-w	.01
+s	-r	+w	.90
+s	-r	-w	.10
-s	+r	+w	.90
-s	+r	-w	.10
-s	-r	+w	0
-s	-r	-w	1.0

Bayesian Network

- Compact representation of the joint distribution
- Conditional independence relationships explicit
- Still represents joint so can be used to answer any probabilistic query

Probabilistic Inference

- Compute probability of a **query** variable (or variables) taking on a value (or set of values) given some **evidence**
- $\Pr[Q \mid E_1=e_1,\ldots,E_k=e_k]$

Using the Joint To Answer Queries

- Joint distribution is sufficient to answer any probabilistic inference question involving variables described in joint
- Can take Bayes Net, construct full joint, and then look up entries where evidence variables take on specified values

But Constructing Joint Expensive & Exact Inference is NP-Hard

- Consider the 3-SAT clause:

$$(x_1 \lor x_2 \lor \neg x_3) \land (\neg x_1 \lor x_3 \lor \neg x_4) \land (x_2 \lor \neg x_2 \lor x_4) \land (\neg x_3 \lor \neg x_4 \lor \neg x_5) \land (x_2 \lor x_5 \lor x_7) \land (x_4 \lor x_5 \lor x_6) \land (\neg x_5 \lor x_6 \lor \neg x_7) \land (\neg x_5 \lor \neg x_6 \lor x_7)$$

which can be encoded by the following Bayes' net:

$P(X_i = 0) = P(X_i = 1) = 0.5$

$Y_1 = X_1 \lor X_2 \lor \neg X_3$

...

$Y_8 = \neg X_5 \lor X_6 \lor X_7$

$Y_{1,2} = Y_1 \land Y_2$

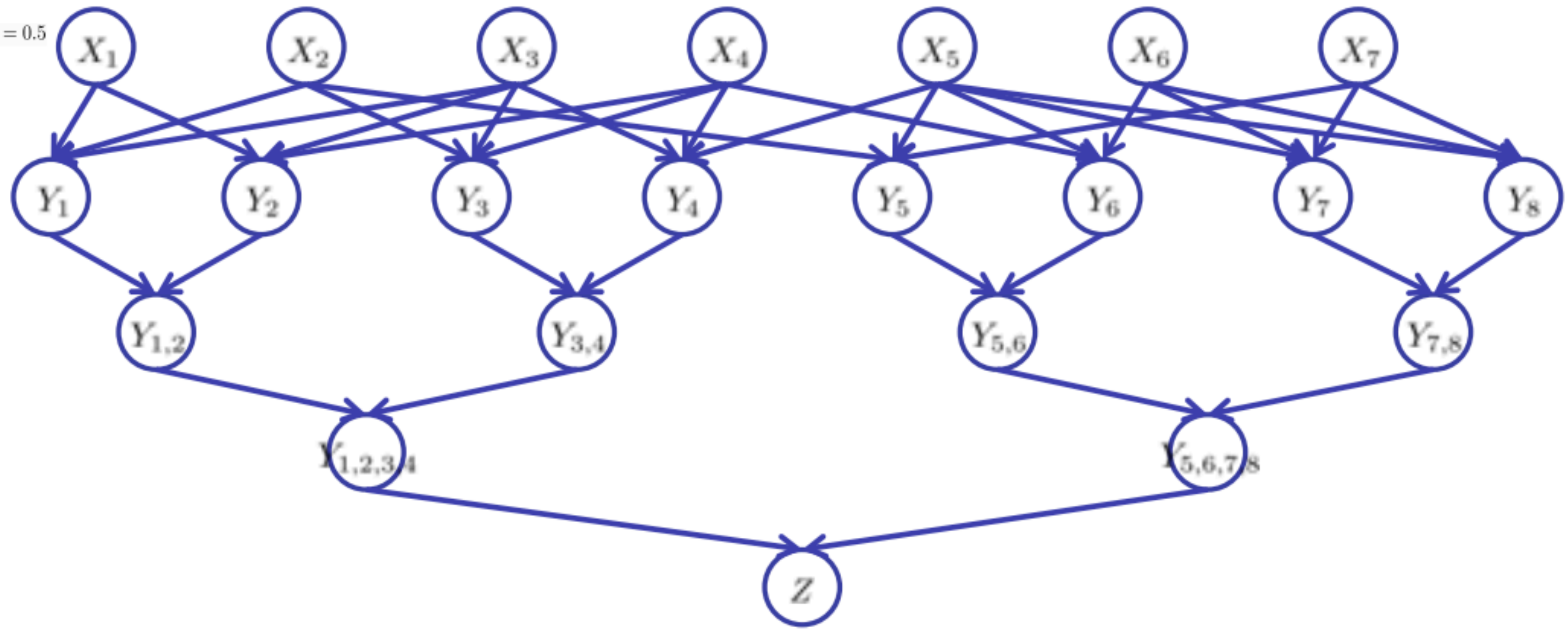
...

$Y_{7,8} = Y_7 \land Y_8$

$Y_{1,2,3,4} = Y_{1,2} \land Y_{3,4}$

$Y_{5,6,7,8} = Y_{5,6} \land Y_{7,8}$

$Z = Y_{1,2,3,4} \land Y_{5,6,7,8}$



If we can answer P(z) equal to zero or not, we answered whether the 3-SAT problem has a solution.

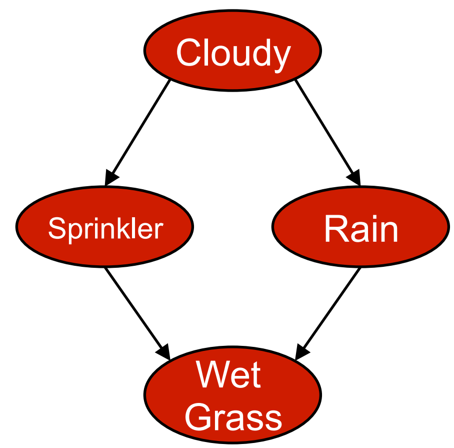
Soln: Approximate Inference

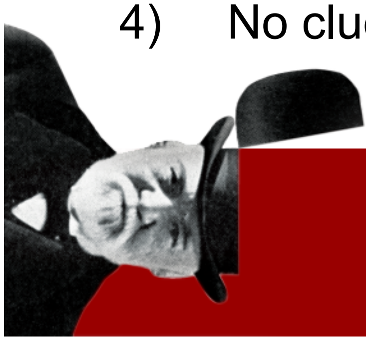
- Use samples to approximate posterior distribution $\Pr[Q \mid E_1=e_1,\ldots,E_k=e_k]$
- Last time
  - Direct sampling
  - Likelihood weighting
- Today
  - Gibbs

Poll: Which Algorithm?

- Evidence: Cloudy=+c, Rain=+r
- Query variable: Sprinkler
- P(Sprinkler|Cloudy=+c,Rain=+r)
- Samples
  - +c,+s,+r,-w
  - +c,-s,-r,-w
  - +c,+s,-r,+w
  - +c,-s,+r,-w
- What algorithm could've generated these samples?

1) Direct sampling
2) Likelihood weighting
3) Both
4) No clue





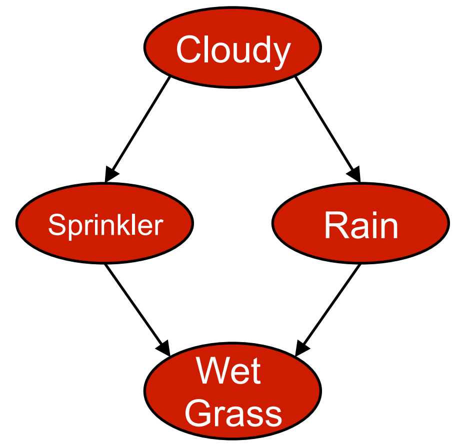
Direct Sampling Recap

Algorithm:

1. Create a topological order of the variables in the Bayes Net

Topological Order

- Any ordering in directed acyclic graph where a node can only appear after all of its ancestors in the graph
- E.g.
  - Cloudy, Sprinkler, Rain, WetGrass
  - Cloudy, Rain, Sprinkler, WetGrass



Direct Sampling Recap

Algorithm:

1. Create a topological order of the variables in the Bayes Net
2. Sample each variable conditioned on the values of its parents
3. Use samples which match evidence variable values to estimate probability of query variable

e.g. P(Sprinkler=+s|Cloudy=+c,Rain=+r) ~ # samples with +s,+c, +r / # samples with +c, +r

- Consistent in limit of infinite samples
- Inefficient (why?)

Consistency

- In the limit of infinite samples, estimated $\Pr[Q \mid E_1=e_1,...,E_k=e_k]$ will converge to true posterior probability
- Desirable property (otherwise always have some error)

Likelihood Weighting Recap

1. Create array TotalWeights
    1. Initialize value of each array element to 0
2. For j=1:N
    1. $w_{tmp} = 1$
    2. Set evidence variables in sample **z**=<$z_1,\ldots z_n$> to observed values
    3. For each variable z_i in topological order
        1. If x_i is an evidence variable
            1. $w_{tmp} = w_{tmp} * P(Z_i = e_i \mid Parents(Z) = \mathbf{x}(Parents(Z_i)))$
        2. Else
            1. Sample x_i conditioned on the values of its parents
    4. Update weight of resulting sample
        1. TotalWeights[**z**]=TotalWeights[**z**]+w_{tmp}

3. Use weights to compute probability of query variable

P(Sprinkler=+s|Cloudy=+c,Rain=+r) ~ $\text{Sum}_{c,r,w}$TotalWeight(+s,c,r,w)/$\text{Sum}_{s,c,r,w}$TotalWeight(s,c,r,w)

LW Consistency

- Probability of getting a sample (**z,e**) where z is a set of values for the non-evidence variables and e is the vals of evidence vars

Sampling distribution for a weighted sample (WS) $\longrightarrow$

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

- Is this the true posterior distribution P(z|e)?
  - No, why?
  - Doesn't consider evidence that is not an ancestor…
  - Weights fix this!

Weighted Probability

- Samples each non-evidence variable z according to

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

- Weight of a sample is

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

- Weighted probability of a sample is

$$\begin{aligned} S_{WS}(\mathbf{z}, \mathbf{e}) w(\mathbf{z}, \mathbf{e}) &= \prod_{i=1}^{l} P(z_i \,|\, parents(Z_i)) \prod_{i=1}^{m} P(e_i \,|\, parents(E_i)) \\ &= P(\mathbf{z}, \mathbf{e}) \end{aligned}$$

From chain rule & conditional indep

Does Likelihood Weighting Produce Consistent Estimates? Yes

$$P(X = x \mid e) = \frac{P(X = x, e)}{P(e)} \propto P(X = x, e)$$

X is query var(s)
E is evidence var(s)
Y is non-query vars

$$\tilde{P}(X = x \mid e) \propto \tilde{P}(X = x, e) = \sum_y N_{WS}(x, y, e) w(x, y, e)$$

# of samples where query variables=x, non-query=y, Evidence=e

$$\approx \sum_y n * S_{WS}(x, y, e) w(x, y, e)$$

as # samples n → infinity

$$= \sum_y P(x, y, e)$$

$$= P(x, e)$$

Example

- When sampling S and R the evidence W=t is ignored
  - Samples with S=f and R=f although evidence rules this out
- Weight makes up for this difference
  - above weight would be 0
- If we have 100 samples with R=t and **total weight 1**, and 400 samples with R=f **and total weight 2**, what is estimate of R=t?
  - = 1/ 3

Limitations of Likelihood Weighting

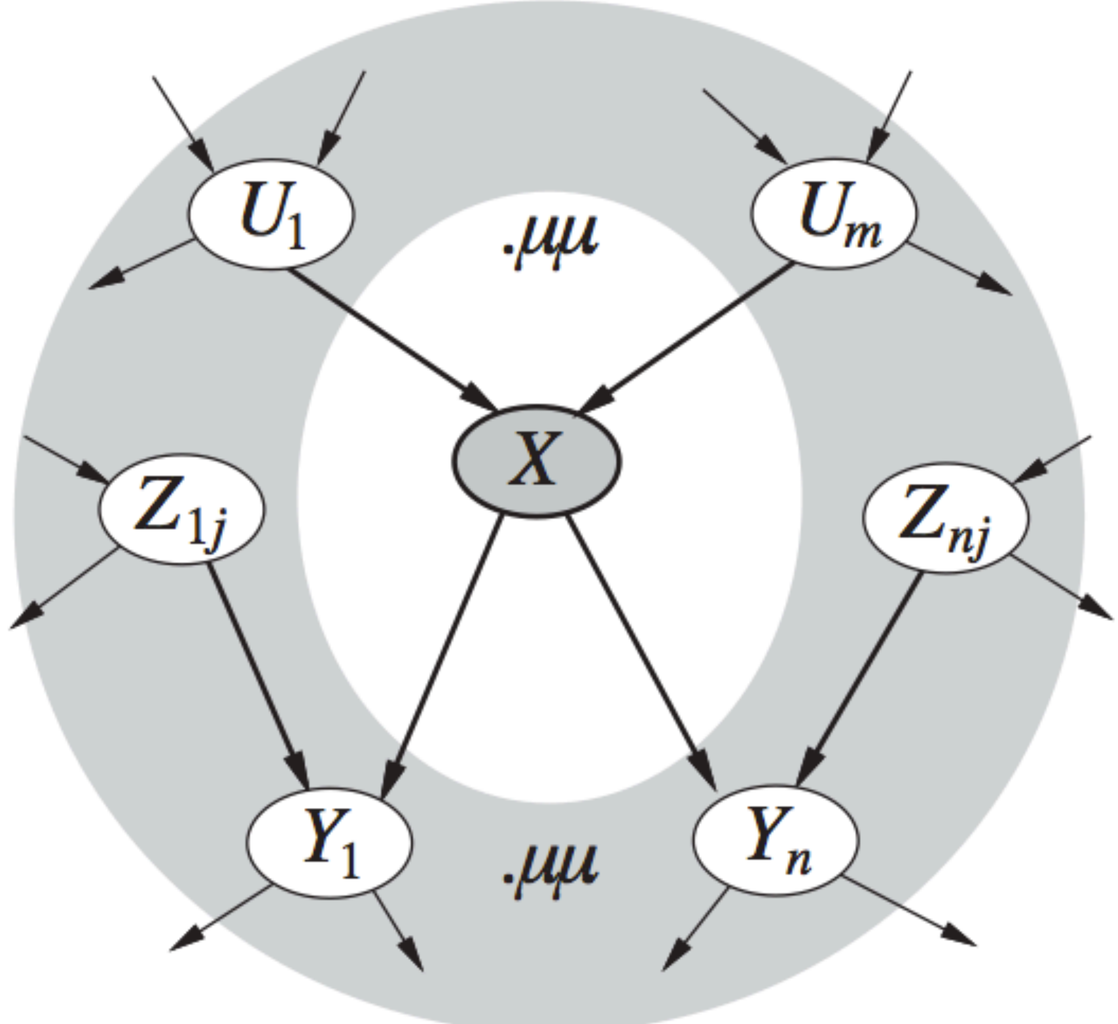
- Poor performance if evidence vars occur later in ordering
- Why?
- Not being used to influence samples!
- Yields samples with low weights

Markov Chain Monte Carlo Methods

- Prior methods generate each new sample from scratch
- MCMC generate each new sample by making a random change to preceding sample
- Can view algorithm as being in a particular state (assignment of values to each variable)

Review: Markov Blanket

- Markov blanket
  - Parents
  - Children
  - Children's parents
- Variable conditionally independent of all other nodes given its Markov Blanket



Gibbs Sampling: Compute P(X|e)

local variables: **N**, a vector of counts for each value of X, initially zero
Z, the nonevidence variables in bn
x, the current state of the network, initially copied from **e**

initialize **x** with random values for the variables in **Z**

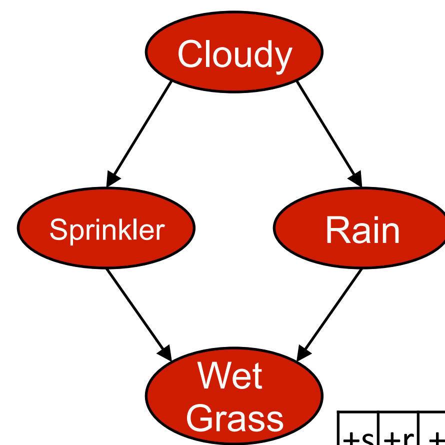
mb(Z_i) = Markov Blanket of Z_i

from Russell & Norvig

Gibbs Sampling Example

- Want Pr(R|S=t,W=t)
- Non-evidence variables are C & R
- Initialize randomly: C= t and R=f
- Initial state (C,S,R,W)= [t,t,f,t]
- Sample C given current values of its Markov Blanket

+c	0.5
-c	0.5



+c	+s	.1
+c	-s	.9
-c	+s	.5
-c	-s	.5

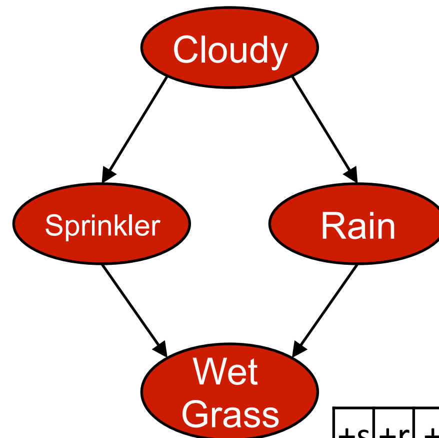
+c	+r	.8
+c	-r	.2
-c	+r	.2
-c	-r	.8

+s	+r	+w	.99
+s	+r	-w	.01
+s	-r	+w	.90
+s	-r	-w	.10
-s	+r	+w	.90
-s	+r	-w	.10
-s	-r	+w	0
-s	-r	-w	1.0

Gibbs Sampling Example

- Want Pr(R|S=t,W=t)
- Non-evidence variables are C & R
- Initialize randomly: C= t and R=f
- Initial state (C,S,R,W)= [t,t,f,t]
- Sample C given current values of its Markov Blanket
- Markov blanket is parents, children and children's parents: for C=S & R
- Sample C given P(C|S=t,R=f)
- First have to compute P(C|S=t,R=f)
- Use exact inference to do this

+c	0.5
-c	0.5



+c	+s	.1
+c	-s	.9
-c	+s	.5
-c	-s	.5

+c	+r	.8
+c	-r	.2
-c	+r	.2
-c	-r	.8

+s	+r	+w	.99
+s	+r	-w	.01
+s	-r	+w	.90
+s	-r	-w	.10
-s	+r	+w	.90
-s	+r	-w	.10
-s	-r	+w	0
-s	-r	-w	1.0

Exercise: compute P(C=t|S=t,R=f)?

- Quick refresher
- Sum rule

$$p(X) = \sum_Y p(X,Y)$$

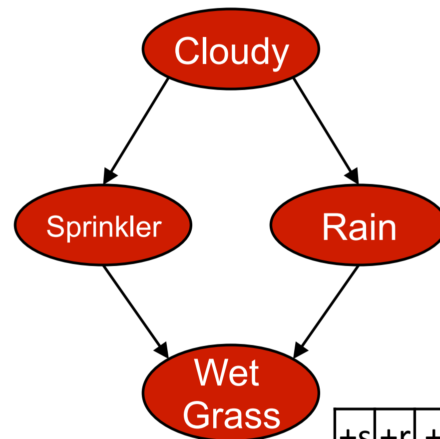
- Product/Chain rule

$$p(X,Y) = p(Y|X)p(X)$$

- Bayes rule

$$p(X|Y) = \frac{p(Y|X)p(X)}{p(Y)}$$

+c	0.5
-c	0.5



+c	+s	.1
+c	-s	.9
-c	+s	.5
-c	-s	.5

+c	+r	.8
+c	-r	.2
-c	+r	.2
-c	-r	.8

+s	+r	+w	.99
+s	+r	-w	.01
+s	-r	+w	.90
+s	-r	-w	.10
-s	+r	+w	.90
-s	+r	-w	.10
-s	-r	+w	0
-s	-r	-w	1.0

Exact Inference Exercise

- P(C|S=t,R=f)
- What is the probability P(C=t | S=t, R= f)?

= P(C=t, S=t, R=f) / (P(S=t,R=f))

Proportional to P(C=t, S=t, R=f)

Use normalization trick, & compute the above for C=t and C=f

P(C=t, S=t, R=f) = P(C=t) P(S=t|C=t) P (R=f | C=t, S=t) product rule

= P(C=t) P(S=t|C=t) P (R=f | C=t) (BN independencies)

= 0.5 * 0.1 * 0.2 = 0.01

P(C=f, S=t, R=f) = P(C=f) P (S=t|C=f) P(R=f|C=f)

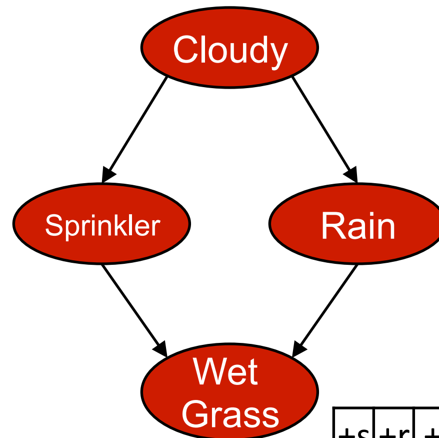
= 0.5 * 0.5 * 0.8 = 0.2

(P(S=t,R=f)) use sum rule = P(C=f, S=t, R=f) + P(C=t, S=t, R=f)

P (C = t | S=t, R= f) = 0.21

P (C=t | S=t, R = f) = 0.01 / 0.21 ~ 0.0476

+c	0.5
-c	0.5



+c	+s	.1
+c	-s	.9
-c	+s	.5
-c	-s	.5

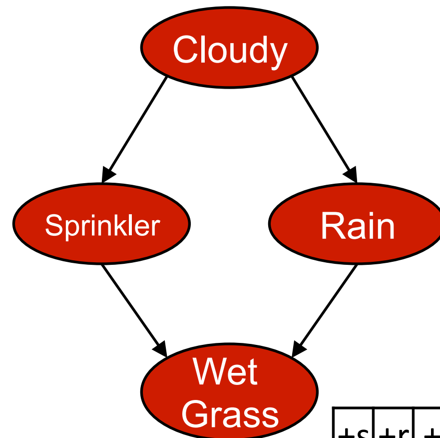
+c	+r	.8
+c	-r	.2
-c	+r	.2
-c	-r	.8

+s	+r	+w	.99
+s	+r	-w	.01
+s	-r	+w	.90
+s	-r	-w	.10
-s	+r	+w	.90
-s	+r	-w	.10
-s	-r	+w	0
-s	-r	-w	1.0

Gibbs Sampling Example

- Want Pr(R|S=t,W=t)
- Non-evidence variables are C & R
- Initialize randomly: C= t and R=f
- Initial state (C,S,R,W)= [t,t,f,t]
- Sample C given current values of its Markov Blanket
- Markov blanket is parents, children and children's parents: for C=S & R
- Exactly compute P(C|S=t,R=f)
- Sample C given P(C|S=t,R=f)
- Get C = f
- New state (f,t,f,t)

+c	0.5
-c	0.5



+c	+s	.1
+c	-s	.9
-c	+s	.5
-c	-s	.5

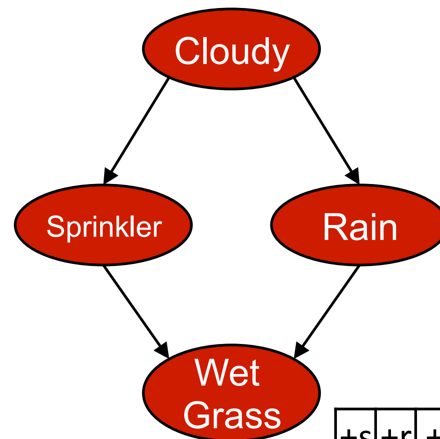
+c	+r	.8
+c	-r	.2
-c	+r	.2
-c	-r	.8

+s	+r	+w	.99
+s	+r	-w	.01
+s	-r	+w	.90
+s	-r	-w	.10
-s	+r	+w	.90
-s	+r	-w	.10
-s	-r	+w	0
-s	-r	-w	1.0

Gibbs Sampling Example

- Want Pr(R|S=t,W=t)
- Initialize non-evidence variables (C and R) randomly to t and f
- Initial state (C,S,R,W)= [t,t,f,t]
- Sample C given current values of its Markov Blanket, p(C|S=t,R=f)
- Suppose result is C=f
- New state (f,t,f,t)
- Sample Rain given its MB
- What is its Markov blanket?

+c	0.5
-c	0.5



+c	+s	.1
+c	-s	.9
-c	+s	.5
-c	-s	.5

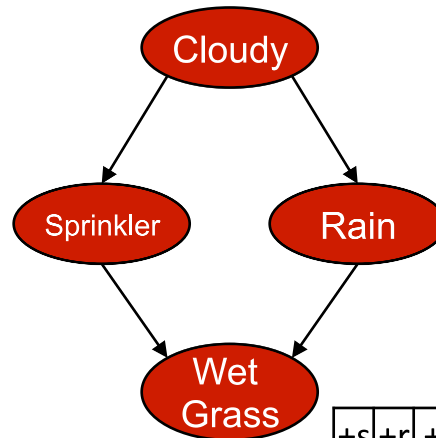
+c	+r	.8
+c	-r	.2
-c	+r	.2
-c	-r	.8

+s	+r	+w	.99
+s	+r	-w	.01
+s	-r	+w	.90
+s	-r	-w	.10
-s	+r	+w	.90
-s	+r	-w	.10
-s	-r	+w	0
-s	-r	-w	1.0

Gibbs Sampling Example

- Want Pr(R|S=t,W=t)
- Initialize non-evidence variables (C and R) randomly to t and f
- Initial state (C,S,R,W)= [t,t,f,t]
- Sample C given current values of its Markov Blanket, p(C|S=t,R=f)
- Suppose result is C=f
- New state (f,t,f,t)
- Sample Rain given its MB, p(R|C=f,S=t,W=t)
- Suppose result is R=t
- New state (f,t,t,t)

+c	0.5
-c	0.5



+c	+s	.1
+c	-s	.9
-c	+s	.5
-c	-s	.5

+c	+r	.8
+c	-r	.2
-c	+r	.2
-c	-r	.8

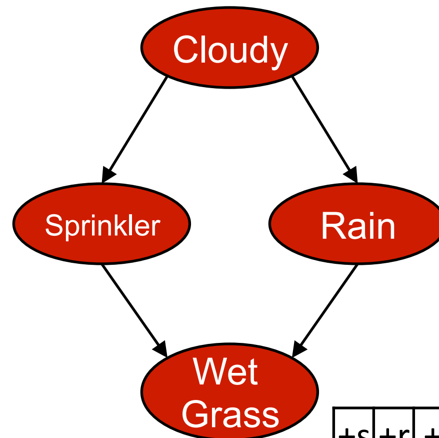
+s	+r	+w	.99
+s	+r	-w	.01
+s	-r	+w	.90
+s	-r	-w	.10
-s	+r	+w	.90
-s	+r	-w	.10
-s	-r	+w	0
-s	-r	-w	1.0

Poll: Gibbs Sampling Ex.

- Want Pr(R|S=t,W=t)
- Initialize non-evidence variables (C and R) randomly to t and f
- Initial state (C,S,R,W)= [t,t,f,t]
- Current state (f,t,t,t)

- What is **not** a possible next state
- 1. (f,t,t,t)
- 2. (t,t,t,t)
- 3. (f,t,f,t)
- **4. (f,f,t,t) (inconsistent w/evid)**
- 5. Not sure

+c	0.5
-c	0.5



+c	+s	.1
+c	-s	.9
-c	+s	.5
-c	-s	.5

+c	+r	.8
+c	-r	.2
-c	+r	.2
-c	-r	.8

+s	+r	+w	.99
+s	+r	-w	.01
+s	-r	+w	.90
+s	-r	-w	.10
-s	+r	+w	.90
-s	+r	-w	.10
-s	-r	+w	0
-s	-r	-w	1.0

Gibbs Sampling

local variables: $\mathbf{N}$, a vector of counts for each value of X, initially zero
$\mathbf{Z}$, the nonevidence variables in bn
$\mathbf{x}$, the current state of the network, initially copied from $\mathbf{e}$

initialize $\mathbf{x}$ with random values for the variables in $\mathbf{Z}$
for $j = 1$ to N **do**
  for each Z_i in $\mathbf{Z}$ **do**
    set the value of Z_i in $\mathbf{x}$ by sampling from $\mathbf{P}(Z_i | mb(Z_i))$
    $\mathbf{N}[x] \leftarrow \mathbf{N}[x] + 1$ where x is the value of X in $\mathbf{x}$
return NORMALIZE($\mathbf{N}$)

This involve inference!

$mb(Z_i)$ = Markov Blanket of Z_i

from Russell & Norvig

Poll

Are Gibbs samples independent?

1. Yes 2. No 3. Not sure

local variables: N, a vector of counts for each value of X, initially zero
Z, the nonevidence variables in bn
x, the current state of the network, initially copied from **e**

initialize **x** with random values for the variables in **Z**
for j = 1 to N **do**
  for each Z_i in **Z do**
    set the value of Z_i in **x** by sampling from $\mathbf{P}(Z_i|mb(Z_i))$
    $\mathbf{N}[x] \leftarrow \mathbf{N}[x] + 1$ where x is the value of X in **x**
return NORMALIZE(**N**)

$mb(Z_i)$ = Markov Blanket of Z_i

from Russell & Norvig

Markov Blanket Sampling

- Want to show $P(Z_i | mb(Z_i))$ is same as $P(Z_i \mid \text{all other variables})$
  - Implies conditional independence of Z_i from rest of network given its Markov Blanket
- Derive equation for computing $P(Z_i | mb(Z_i))$

Probability Given Markov Blanket

$$P(x_i' \mid mb(X_i)) = \alpha \, P(x_i' \mid parents(X_i)) \times \prod_{Y_j \in Children(X_i)} P(y_j \mid parents(Y_j))$$

Why is Gibbs Consistent?

- Sampling process settles into a stationary distribution where long-term fraction of time spent in each state is exactly equal to posterior probability
  - → Implies that if draw enough samples from this stationary distribution, will get consistent estimate because sampling from true posterior

Markov Chain

- Let P(**x** → **x**’) be probability the sampling process makes a transition from **x** (some state) to **x’** (some other state)
  - E.g. (t,t,f,t) → (t,f,f,t)
- Run sampling for t steps
- $P_t(\mathbf{x})$ is probability system is in state **x** at time t
- Next state $P_{t+1}(\mathbf{x'}) = \mathrm{Sum}_{\mathbf{x}}\ P_t(\mathbf{x})\ P(\mathbf{x} \rightarrow \mathbf{x'})$

Stationary Distribution

- Let $P(\mathbf{x} \rightarrow \mathbf{x}')$ be probability the process makes a transition from $\mathbf{x}$ to $\mathbf{x}'$
- $P_t(\mathbf{x})$ is probability system is in state $\mathbf{x}$ at time t
- $P_{t+1}(\mathbf{x}') = \mathrm{Sum}_{\mathbf{x}}\, P_t(\mathbf{x})\, P(\mathbf{x} \rightarrow \mathbf{x}')$
- Reached stationary distribution if $P_{t+1}(\mathbf{x}')=P_t(\mathbf{x})$
- Call stationary distribution π
  - Must satisfy π(**x’**) = \sum_{**x**} π(**x**) P(**x** → **x**’) for all **x’**
- If $P(\mathbf{x} \rightarrow \mathbf{x}')$ is ergodic, exactly one such π for any given $P(\mathbf{x} \rightarrow \mathbf{x}')$

Detailed Balance

- Let P(**x** → **x**') be probability the process makes a transition from **x** to **x'**
- P_t(**x**) is probability system is in state **x** at time t
- Stationary distribution π
  - Satisfies π(**x'**) = \sum_{**x**} π(**x**) P(**x** → **x**') for all **x'**
- Detailed balance: inflow = outflow
- π(**x**) P(**x** → **x**') = π(**x'**) P(**x'** → **x**) for all **x, x'**

$$\sum_x \pi(x')P(x->x') = \sum_x \pi(x')P(x'->x) = \pi(x')$$

- Proof on board

Proving Gibbs Samples from True Posterior

- General Gibbs: sample the value of a new variable conditioned on all the other variables
- Can prove this version of Gibbs satisfies detailed balance equation with stationary distribution of P(X|e)
- Then use prior result that sampling conditioned on all variables is equivalent to sampling given Markov Blanket for Bayes Nets
- See text for recap

Gibbs Sampling

- Samples are valid once reach stationary distribution
- When do we reach stationary distribution?
    - Unclear…

What You Should Know

- Define probabilistic inference
- How to define a Bayes Net given a real example
- How joint can be used to answer any query
- Complexity of exact inference
- Approximation inference (direct, likelihood, Gibbs)
  - Be able to implement and run algorithm
  - Compare benefits and limitations of each