Monte Carlo Markov Chain 1

1

1

MCMC

Limitations of LW:

- Evidence affects sampling only for nodes that are its descendants
- For nondescendants, the weights account for the effect of the evidence
- If evidence is at the leaves, we are sampling from the prior distribution (and not the posterior which is what we want)

2

2

MCMC

Strategy used by MCMC

- Generate a <u>sequence</u> of samples
- Initial samples generated from the prior
- Successive samples generated progressively closer to the posterior

Applies to both directed and undirected models. We'll use a distribution P_Φ defined in terms of a set of factors Φ

3

3

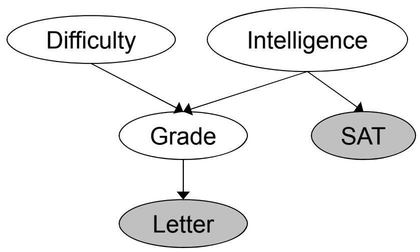
Gibbs Sampling

4

4

Gibbs Sampling

Example: Suppose we have as evidence *SAT = High and Letter = Weak* (nodes are shaded grey)



<u>Factors</u>:

- P(I)
- P(D)
- P(G | I,D)

<u>Reduced Factors</u>:

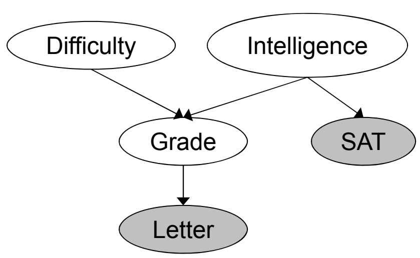
- P(S=high | I)
- P(L=weak | G)

Eliminate all rows that are inconsistent with the evidence in all factors (see pg 111 of textbook)

5

5

Gibbs Sampling



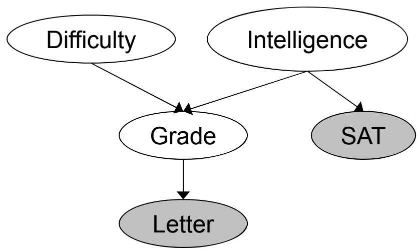
Start with an initial sample eg: $\boldsymbol{x}^{(0)}$ = *(D = high, I = low, G = B, S = high, L = weak)*

- *D*, *I* and *G* could be set in any way, for instance by forward sampling, to get $D^{(0)}$ = *high*, $I^{(0)}$ = *low*, $G^{(0)}$ = *B*
- *S=high* and *L=weak* are observed

6

6

Gibbs Sampling



Resample non-evidence nodes, one at a time, in some order eg. G, I, D.

If we sample X_i, keep other nodes clamped at the values of the current state *(D = high, I = low, G = B, S = high, L = weak)*

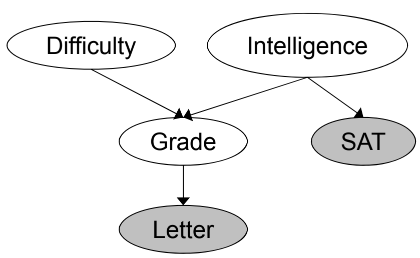
To sample $G^{(1)}$, we compute $P_\Phi(G \mid D=high, I=low, S=high, L=weak)$:

$$P_\Phi(G|D = high, I = low, S = high, L = weak)$$
$$= \frac{P(I = high)P(D = low)P(G|I = low, D = High)P(L = low|G)P(S = high|I = low)}{\sum_G P(I = high)P(D = low)P(G|I = low, D = High)P(L = low|G)P(S = high|I = low)}$$
$$= \frac{P(G|I = low, D = high)P(Letter = weak|G)}{\sum_G P(G|I = low, D = high)P(Letter = weak|G)}$$

7

7

Gibbs Sampling



- Suppose we obtain $G^{(1)}$ = C.
- Now sample $I^{(1)}$ from P_Φ(I | D=high, G=C, S=high, L=weak). Note it is conditioned on $G^{(1)}$=C
- Say we get $I^{(1)}$=high
- Now sample $D^{(1)}$ from P_Φ(D | G=C, I = high, S=high, L=weak). Say you get $D^{(1)}$ = high
- The first iteration of sampling produces $\mathbf{x}^{(1)}$ = (I = high, D = high, G = C, S=high, L=weak)
- Iterate...

8

8

Gibbs Sampling

- $P_\Phi(G \mid D{=}high, I{=}low, S{=}high, L{=}weak)$ takes downstream evidence $L{=}weak$ into account (makes it closer to the posterior distribution $P(X \mid \mathbf{e})$)
- Early on, $P_\Phi(G \mid D{=}high, I{=}low, S{=}high, L{=}weak)$ very much like the prior $P(X)$ because it uses values for I and D sampled from $P(X)$
- On next iteration, resampling I and D conditioned on new value of G brings the sampling distribution closer to the posterior
- Sampling distribution gets progressively closer and closer to the posterior

Gibbs Sampling

```
Procedure Gibbs-Sample (
   X            // Set of variables to be sampled
   Φ            // Set of factors defining P_Φ
   P^(0)(X),    // Initial state distribution
   T            // Number of time steps
)
Sample x^(0) from P^(0)(X)
for t=1, ..., T
  x^(t) ← x^(t-1)
  for each X_i ∈ X
    Sample x_i^(t) from P_Φ(X_i | x_-i)
    // Change X_i in x^(t)
return x^(0), ..., x^(T)
```

Gibbs Sampling

Gibbs sampling with evidence

- Reduce all factors by the observations **e**
- The distribution P_Φ corresponds to $P(X|\mathbf{e})$

11

11

Markov Chains

12

12

Markov Chains

- (Informally) A Markov chain is a graph of states over which the sampling algorithm takes a random walk
- Note: the graph is not the graphical model but a graph over the possible assignments to a set of variables ***X***

13

13

Markov Chains

- A Markov chain is defined via a state space *Val*(***X***) and a model that defines, for every state $x \in Val(\mathbf{X})$ a next-state distribution over *Val*(***X***).
- More precisely, the transition model $\mathcal{T}$ specifies for each pair of states x, x' the probability $\mathcal{T}(x \rightarrow x')$ of going from x to x'.
- A homogeneous Markov chain is one where the system dynamics do not change over time

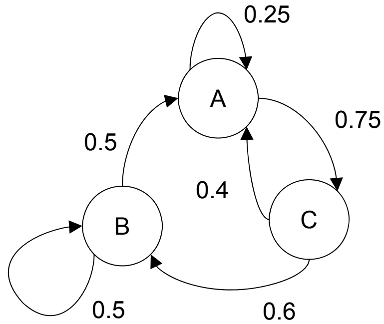
14

14

Markov Chains

Example of a Markov Chain with Val(X)={A,B,C}:

State Transition Diagram View



Conditional Probability Distribution View

X_{t-1}	X_t	$P(X_t \mid X_{t-1})$
A	A	0.25
A	B	0
A	C	0.75
B	A	0.5
B	B	0.5
B	C	0
C	A	0.4
C	B	0.6
C	C	0

Markov Chains

- Random sampling process defines a random sequence of states $\mathbf{x}^{(0)}$, $\mathbf{x}^{(1)}$, $\mathbf{x}^{(2)}$, …
- $\mathbf{X}^{(t)}$ is a random variable:
- Need initial state distribution $P^{(0)}(\mathbf{X}^{(0)})$
- Probability that next state is **x**' can be computed as:

$$P^{(t+1)}(X^{(t+1)} = x') = \underbrace{\sum_{x \in Val(X)} P^{(t)}(X^{(t)} = x)}\underbrace{\mathcal{T}(x \rightarrow x')}$$

Sum over all states that the chain could have been at time t

Probability of transition from x to x'

Markov Chains

How to generate a Markov Change Monte Carlo trajectory:

```
Procedure MCMC-Sample (
    P^(0)(X),  // Initial state distribution
    𝒯,         // Markov chain transition model
    T          // Number of time steps
)
1.  Sample x^(0) from P^(0)(X)
2.  for t = 1, …, T
3.      Sample x^(t) from 𝒯(x^(t-1) → X)
4.  return x^(0), …, x^(T)
```

The big question: does $P^{(t)}$ converge and what to?

17

17

Markov Chains

- When the process converges, we expect:

$$P^{(t)}(\boldsymbol{x}') \approx P^{(t+1)}(\boldsymbol{x}') = \sum_{\boldsymbol{x} \in Val(\boldsymbol{X})} P^{(t)}(\boldsymbol{x}) \mathcal{T}(\boldsymbol{x} \rightarrow \boldsymbol{x}')$$

- A distribution $\pi(\mathbf{X})$ is a stationary distribution for a Markov chain $\mathcal{T}$ if it satisfies:

$$\pi(\boldsymbol{X} = \boldsymbol{x}') = \sum_{\boldsymbol{x} \in Val(\boldsymbol{X})} \pi(\boldsymbol{X} = \boldsymbol{x}) \mathcal{T}(\boldsymbol{x} \rightarrow \boldsymbol{x}')$$

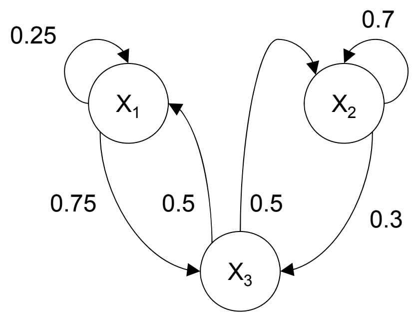
- A stationary distribution is also called an invariant distribution

18

18

Markov Chains

Another example:



To find the stationary distribution:

$\pi(x_1) = 0.25\pi(x_1)+0.5\pi(x_3)$

$\pi(x_2) = 0.7\pi(x_2)+0.5\pi(x_3)$

$\pi(x_3) = 0.75\pi(x_1)+0.3\pi(x_2)$

$\pi(x_1) + \pi(x_2) + \pi(x_3) = 1$

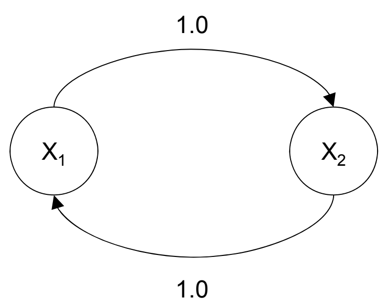
Solving these simultaneous equations gives: $\pi(x_1) = 0.2$, $\pi(x_2) = 0.5$, $\pi(x_3) = 0.3$

19

19

Markov Chains

- Bad news: no guarantee that MCMC sampling process converges to a stationary distribution
- Example of a periodic Markov chain (periodic = fixed cyclic behavior)
  - Start with $P^{(0)}(x_1) = 1$
  - $P^{(t)}(x_1) = 1$ if t is even
  - $P^{(t)}(x_2) = 1$ if t is odd



20

20

Markov Chains

- No guarantee that stationary distribution is unique – depends on $P^{(0)}$
  - This happens if the chain is reducible: has states that are not reachable from each other
- We will restrict our attention to Markov chains that have a stationary distribution which is reached from any starting distribution $P^{(0)}$

21

21

Markov Chains

- To meet this restriction, we need the chain to be regular
- A Markov chain is said to be regular if there exists some number k such that, for every x, $x' \in Val(\boldsymbol{X})$, the probability of getting from x to x' in exactly k steps is > 0
- Theorem 12.3: If a finite state Markov chain $\mathcal{T}$ is regular, then it has a unique stationary distribution

22

22

Markov Chains

- Define $\mathcal{T}_i$ to be a transition model called a kernel
- For graphical models, define a kernel $\mathcal{T}_i$ for each variable $X_i \in \boldsymbol{X}$
- Define $\mathbf{X}_{-i} = \mathcal{X} - \{X_i\}$ and let $\boldsymbol{x}_i$ denote an instantiation to $\boldsymbol{X}_i$
- The model $\mathcal{T}_i$ takes a state $(\boldsymbol{x}_{-i}, x_i)$ and transitions to a state $(\boldsymbol{x}_{-i}, x_i')$

23

23

Gibbs Sampling Revisited

24

24

Gibbs Sampling Revisited

How do we use MCMC on a graphical model?

- Want to generate samples from the posterior $P(\boldsymbol{X}|\boldsymbol{E}=\boldsymbol{e})$ where $\boldsymbol{X}=\mathcal{X} - \boldsymbol{E}$
- Define a chain where $P(\boldsymbol{X}|\boldsymbol{e})$ is the stationary distribution
- States are instantiations $\boldsymbol{x}$ to $\mathcal{X} - \boldsymbol{E}$
- Need transition function that converges to stationary distribution $P(\boldsymbol{X}|\boldsymbol{e})$
- For convenience: define $P_\Phi = P(\boldsymbol{X}|\boldsymbol{e})$ where the factors in Φ are reduced by the evidence $\boldsymbol{e}$

25

25

Gibbs Sampling Revisited

Using the MCMC framework, the transition model for Gibbs Sampling is:

$$\mathcal{T}_i((\boldsymbol{x}_{-i}, x_i) \rightarrow (\boldsymbol{x}_{-i}, x_i')) = P(x_i' \mid \boldsymbol{x}_{-i})$$

And the posterior distribution $P_\Phi(\boldsymbol{X}) = P(\mathcal{X}|\boldsymbol{e})$ is a stationary distribution of this process

26

26

Gibbs Sampling Revisited

Gibbs sampling on a Bayesian network is efficient
Note: $Pa(x_i)$ = Parents of x_i, $Ch(x_i)$ = Children of x_i

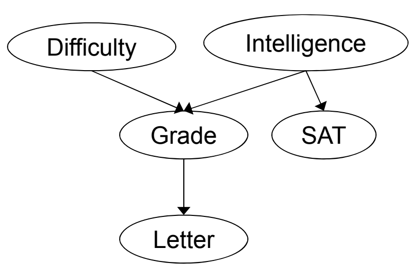
$$P(X_i|x_1,\dots,x_{i-1},x_{i+1},x_n)$$
$$=\frac{P(x_1,\dots,x_{i-1},x_i,x_{i+1},x_n)}{P(x_1,\dots,x_{i-1},x_{i+1},x_n)}=\frac{P(x_1,\dots,x_{i-1},x_i,x_{i+1},x_n)}{\sum_{x_i}P(x_1,\dots,x_{i-1},x_i,x_{i+1},x_n)}$$
$$=\frac{\prod_{j=1}^{n}P\big(x_j\big|Pa(x_j)\big)}{\sum_{x_i}\prod_{j=1}^{n}P\big(x_j\big|Pa(x_j)\big)}$$
$$=\frac{\prod_{x_j\notin\{x_i,Ch(x_i)\}}P\big(x_j\big|Pa(x_j)\big)P(x_i|Pa(x_i))\prod_{x_{k\in Ch(x_i)}}P(x_k|x_i,\mathit{other\ parents})}{\prod_{x_j\notin\{x_i,Ch(x_i)\}}P\big(x_j\big|Pa(x_j)\big)\sum_{x_i}P(x_i|Pa(x_i))\prod_{x_{k\in Ch(x_i)}}P(x_k|x_i,\mathit{other\ parents})}$$
$$=\frac{P(x_i|Pa(x_i))\prod_{x_{k\in Ch(x_i)}}P(x_k|x_i,\mathit{other\ parents})}{\sum_{x_i}P(x_i|Pa(x_i))\prod_{x_{k\in Ch(x_i)}}P(x_k|x_i,\mathit{other\ parents})}$$

Depends only on the
CPDs of X_i and its children

27

27

Gibbs Sampling Revisited



Student Example revisited:

Define:

$\mathcal{T}$((I,G,D,S=high,L=weak) → (I', G, D, S=high, L=weak)) = P(I|G,D,S=high,L=weak)

Sample from the distribution below:

$$P(I'|G,D,S=high,L=weak)$$
$$=\frac{P(I')P(G|I',D)P(S=high|I')}{\sum_{I''}P(I''=i'')P(G|I''=i'',D)P(S=high|I''=i'')}$$

28

28

Gibbs Sampling Revisited

Block Gibbs Sampling

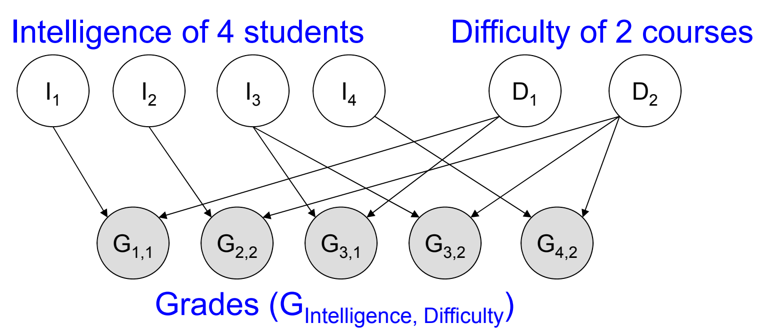
- Can sample more than a single variable X_i at a time
- Partition $\boldsymbol{X}$ into disjoint blocks of variables $\boldsymbol{X_1}, \ldots, \boldsymbol{X_k}$
- Then sample $P_{\Phi}(\boldsymbol{X_i} \mid \boldsymbol{X_1}=\boldsymbol{x_1}, \ldots, \boldsymbol{X_{i-1}}=\boldsymbol{x_{i-1}}, \boldsymbol{X_{i+1}}=\boldsymbol{x_{i+1}}, \ldots, \boldsymbol{X_k}=\boldsymbol{x_k})$
- Takes longer range transitions

29

29

Gibbs Sampling Revisited

Example of Block Gibbs Sampling



- Step t: Sample all of the I variables as a block, given Ds and Gs (since Is are conditionally independent from each other given Ds)
- Step t+1: Sample all of the D variables as a block, given Is and Gs (since Ds are conditionally independent of each other given Is)

30

30

Gibbs Sampling Revisited

Need to compute $P_{\Phi}(\boldsymbol{X}_i \mid \boldsymbol{X}_1=\boldsymbol{x}_1, ..., \boldsymbol{X}_{i-1}=\boldsymbol{x}_{i-1}, \boldsymbol{X}_{i+1}=\boldsymbol{x}_{i+1}, \boldsymbol{X}_k = \boldsymbol{x}_k)$

- Efficient if variables in each block (eg. **I**) are independent given the variables outside the block (eg. **D**)
- In general, full independence is not essential – need some sort of structure to the block-conditional distribution

31

31

Gibbs Sampling Revisited

- Gibbs chain not necessarily regular and may not converge to a unique stationary distribution
- Only guaranteed to be regular if $P(X_i \mid \boldsymbol{X}_{-i})$ is positive for every value of X_i
- Theorem 12.4: Let $\mathcal{H}$ be a Markov network such that all of the clique potentials are strictly positive. Then the Gibbs-sampling Markov chain is regular.

32

32