CS 6347

Lecture 11

MCMC Sampling Methods

Beyond Monte Carlo Methods

- All of the methods discussed so far can have serious limitations depending on the quantity being estimated
- Idea: instead of having a single proposal distribution, why not have an adaptive proposal distribution that depends on the previous sample?

$q(x|x')$ where x' is the previous sample and x is the new assignment to be sampled

Markov Chains

- A Markov chain is a sequence of random variables $X_1, \ldots, X_n \in S$ such that

$$p(x_{n+1}|x_1, \ldots, x_n) = p(x_{n+1}|x_n)$$

- The set S is called the state space, and $p(X_{n+1} = b|X_n = a)$ is the probability of transitioning from state a to state b at step n

- As a Bayesian network or a MRF, the joint distribution over the first n steps factorizes over a chain

Markov Chains

- When the probability of transitioning between two states does not depend on time, we call it a time homogeneous Markov chain
  - Represent it by a $|S| \times |S|$ transition matrix P
    - $P_{ij} = p(X_{n+1} = j | X_n = i)$
    - P is a **stochastic** matrix (all rows sum to one)
  - Draw it as a directed graph over the state space with an arrow from $a \in S$ to $b \in S$ labelled by the probability of transitioning from a to b

Markov Chains

- Given some initial distribution over states $p(x_1)$
  - Represent $p(x_1)$ as a length $|S|$ vector, π_1
  - The probability distribution after n steps is given by

$$\pi_n = \pi_1 P^n$$

- Typically interested in the long term (i.e., what is the state of the system when n is large)
- In particular, we are interested in steady-state distributions μ such that $\mu = \mu P$
  - A given chain may or may not converge to a steady state

Markov Chains

- Theorem: every **irreducible, aperiodic** Markov chain converges to a unique steady state distribution independent of the initial distribution
  - Irreducible: the directed graph of transitions is strongly connected
  - Aperiodic: $p(X_n = i \mid X_1 = i) > 0$ for all large enough n
- If the state graph is strongly connected and there is a non-zero probability of remaining in any state, then the chain is irreducible and aperiodic

MCMC Sampling

- Markov chain Monte Carlo sampling
  - Construct a Markov chain where the stationary distribution is the one we want to sample from
  - Use the Markov chain to generate samples from the distribution
  - Use the same Monte Carlo estimation strategy as before
  - Will let us sample conditional distributions easily as well!

Gibbs Sampling

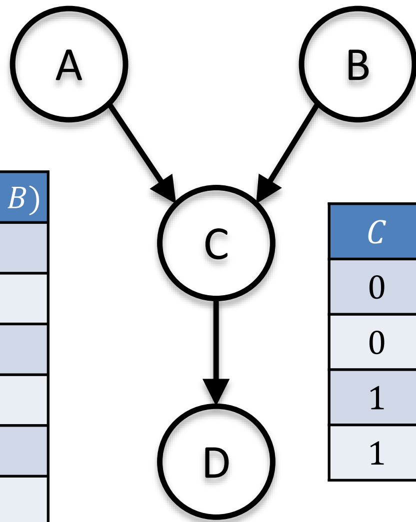
- Let's consider a MRF with $p(x) = \frac{1}{Z}\prod_C \psi_C(x_C)$
- Choose an initial assignment x
- Fix an ordering of the variables (any order is fine)
- For each $j \in V$ in order
  – Draw a sample z from $p(X_j|x_{V\setminus \mathrm{j}})$ using the current x
  – Set $x_j = z$
- Repeat

Gibbs Sampling

- Given that $p(x) = \frac{1}{Z}\prod_C \psi_C(x_C)$ we can sample from $p(X_j|x_{N(j)})$
  - First sampling algorithm that actually lets us exploit the graph structure
  - For Bayesian networks, it reduces to $p(X_j|x_{MB(j)})$ where $MB(j)$ is j's Markov blanket (j's parents, children, and its childrens' parents)

Gibbs Sampling

A	$P(A)$
0	.3
1	.7



B	$P(B)$
0	.4
1	.6

Order: A, B, C, D, A, B, C, D, …

A	B	C	$P(C\|A,B)$
0	0	0	.1
0	0	1	.9
0	1	0	.2
0	1	1	.8
1	0	0	.01
1	0	1	.99
1	1	0	.25
1	1	1	.75

C	D	$P(D\|C)$
0	0	.3
0	1	.7
1	0	.4
1	1	.6

A	B	C	D
0	0	0	0

(1) Sample from $p(x_A | x_B = 0, x_C = 0, x_D = 0)$

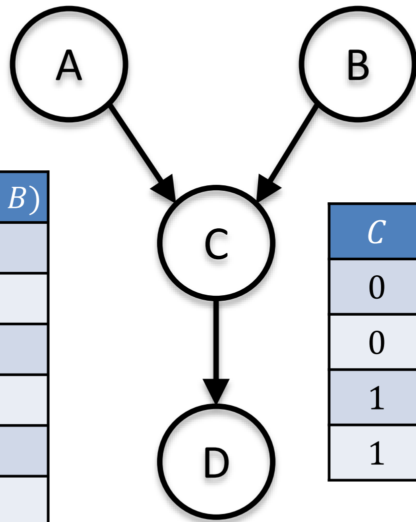
Using Bayes rule, $p(x_A | x_B = 0, x_C = 0) \propto p(x_A) p(x_C = 0 | x_A, x_B = 0)$

$p(x_A = 0 | x_B = 0, x_C = 0) \propto .3 \cdot .1 = \; .03$

$p(x_A = 1 | x_B = 0, x_C = 0) \propto .7 \cdot .01 = .007$

Gibbs Sampling

A	$P(A)$
0	.3
1	.7



B	$P(B)$
0	.4
1	.6

Order: A, B, C, D, A, B, C, D, …

A	B	C	$P(C\|A,B)$
0	0	0	.1
0	0	1	.9
0	1	0	.2
0	1	1	.8
1	0	0	.01
1	0	1	.99
1	1	0	.25
1	1	1	.75

C	D	$P(D\|C)$
0	0	.3
0	1	.7
1	0	.4
1	1	.6

A	B	C	D
0	0	0	0
0			

(1) Sample from $p(x_A|x_B = 0, x_C = 0, x_D = 0)$

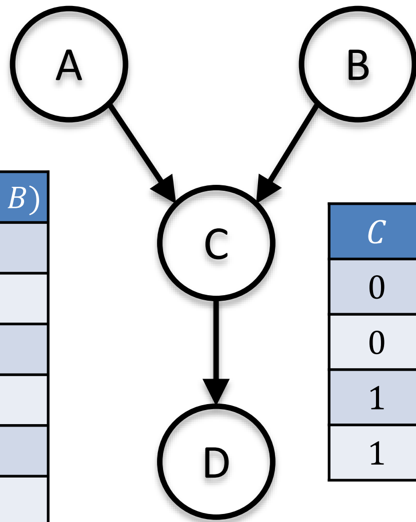
Using Bayes rule, $p(x_A|x_B = 0, x_C = 0) \propto p(x_A)p(x_C = 0|x_A, x_B = 0)$

$p(x_A = 0|x_B = 0, x_C = 0) \propto .3 \cdot .1 \rightarrow .811$

$p(x_A = 1|x_B = 0, x_C = 0) \propto .7 \cdot .01 \rightarrow .189$

Gibbs Sampling

A	$P(A)$
0	.3
1	.7



B	$P(B)$
0	.4
1	.6

Order: A, B, C, D, A, B, C, D, …

A	B	C	$P(C\|A,B)$
0	0	0	.1
0	0	1	.9
0	1	0	.2
0	1	1	.8
1	0	0	.01
1	0	1	.99
1	1	0	.25
1	1	1	.75

C	D	$P(D\|C)$
0	0	.3
0	1	.7
1	0	.4
1	1	.6

A	B	C	D
0	0	0	0
0			

(1) Sample from $p(x_B|x_A = 0, x_C = 0, x_D = 0)$

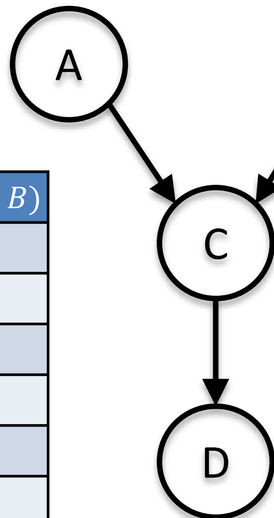
Using Bayes rule, $p(x_B|x_A = 0, x_C = 0) \propto p(x_B)p(x_C = 0|x_A, x_B = 0)$

$p(x_B = 0|x_A = 0, x_C = 0) \propto .4 \cdot .1 = .04$

$p(x_B = 1|x_A = 0, x_C = 0) \propto .6 \cdot .2 = .12$

Gibbs Sampling

A	$P(A)$
0	.3
1	.7



B	$P(B)$
0	.4
1	.6

Order: A, B, C, D, A, B, C, D, …

A	B	C	$P(C\|A,B)$
0	0	0	.1
0	0	1	.9
0	1	0	.2
0	1	1	.8
1	0	0	.01
1	0	1	.99
1	1	0	.25
1	1	1	.75

C	D	$P(D\|C)$
0	0	.3
0	1	.7
1	0	.4
1	1	.6

A	B	C	D
0	0	0	0
0	1		

(1) Sample from $p(x_B|x_A = 0, x_C = 0, x_D = 0)$

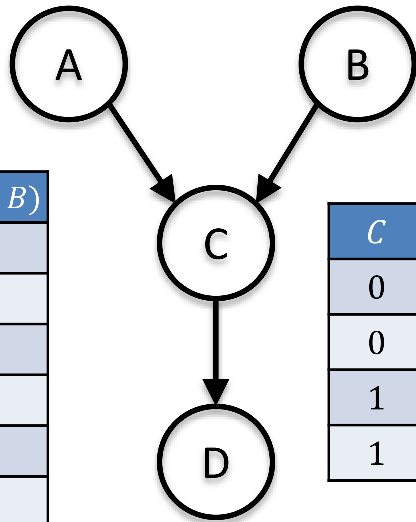
Using Bayes rule, $p(x_B|x_A = 0, x_C = 0) \propto p(x_B)p(x_C = 0|x_A, x_B = 0)$

$p(x_B = 0|x_A = 0, x_C = 0) \propto .4 \cdot .1 \rightarrow .25$

$p(x_B = 1|x_A = 0, x_C = 0) \propto .6 \cdot .2 \rightarrow .75$

Gibbs Sampling

A	$P(A)$
0	.3
1	.7



B	$P(B)$
0	.4
1	.6

Order: A, B, C, D, A, B, C, D, …

A	B	C	$P(C\|A,B)$
0	0	0	.1
0	0	1	.9
0	1	0	.2
0	1	1	.8
1	0	0	.01
1	0	1	.99
1	1	0	.25
1	1	1	.75

C	D	$P(D\|C)$
0	0	.3
0	1	.7
1	0	.4
1	1	.6

A	B	C	D
0	0	0	0
0	1		

(1) Sample from $p(x_C | x_A = 0, x_B = 1, x_D = 0)$

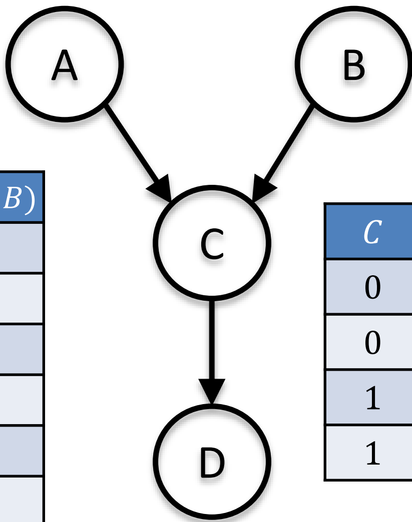
Using Bayes rule, $p(x_C | x_A = 0, x_B = 1, x_D = 0) \propto p(x_C | x_A = 0, x_B = 1) p(x_D = 0 | x_C)$

$p(x_C = 0 | x_A = 0, x_B = 1, x_D = 0) \propto .2 \cdot .3 = .06$

$p(x_C = 1 | x_A = 0, x_B = 1, x_D = 0) \propto .8 \cdot .4 = .32$

Gibbs Sampling

A	$P(A)$
0	.3
1	.7



B	$P(B)$
0	.4
1	.6

Order: A, B, C, D, A, B, C, D, …

A	B	C	$P(C\|A,B)$
0	0	0	.1
0	0	1	.9
0	1	0	.2
0	1	1	.8
1	0	0	.01
1	0	1	.99
1	1	0	.25
1	1	1	.75

C	D	$P(D\|C)$
0	0	.3
0	1	.7
1	0	.4
1	1	.6

A	B	C	D
0	0	0	0
0	1	1	

(1) Sample from $p(x_C|x_A = 0, x_B = 1, x_D = 0)$

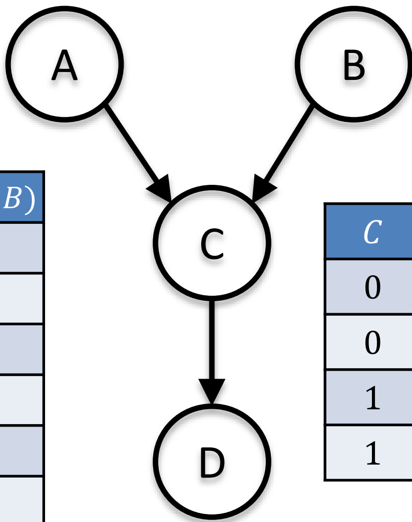
Using Bayes rule, $p(x_C|x_A = 0, x_B = 1, x_D = 0) \propto p(x_C|x_A = 0, x_B = 1)p(x_D = 0|x_C)$

$p(x_C = 0|x_A = 0, x_B = 1, x_D = 0) \propto .2 \cdot .3 \rightarrow .158$

$p(x_C = 1|x_A = 0, x_B = 1, x_D = 0) \propto .8 \cdot .4 \rightarrow .842$

Gibbs Sampling

A	$P(A)$
0	.3
1	.7



B	$P(B)$
0	.4
1	.6

Order: A, B, C, D, A, B, C, D, …

A	B	C	$P(C\|A,B)$
0	0	0	.1
0	0	1	.9
0	1	0	.2
0	1	1	.8
1	0	0	.01
1	0	1	.99
1	1	0	.25
1	1	1	.75

C	D	$P(D\|C)$
0	0	.3
0	1	.7
1	0	.4
1	1	.6

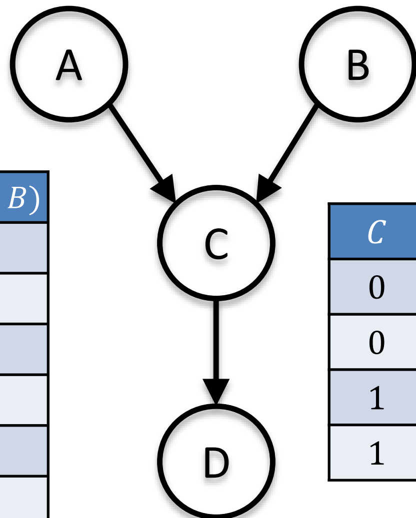
A	B	C	D
0	0	0	0
0	1	1	

(1) Sample from $p(x_D |\, x_C = 1)$

$$p(x_D = 0|x_C = 1) = .4$$
$$p(x_D = 1|x_C = 1) = .6$$

Gibbs Sampling

A	$P(A)$
0	.3
1	.7



B	$P(B)$
0	.4
1	.6

Order: A, B, C, D, A, B, C, D, …

A	B	C	$P(C\|A,B)$
0	0	0	.1
0	0	1	.9
0	1	0	.2
0	1	1	.8
1	0	0	.01
1	0	1	.99
1	1	0	.25
1	1	1	.75

C	D	$P(D\|C)$
0	0	.3
0	1	.7
1	0	.4
1	1	.6

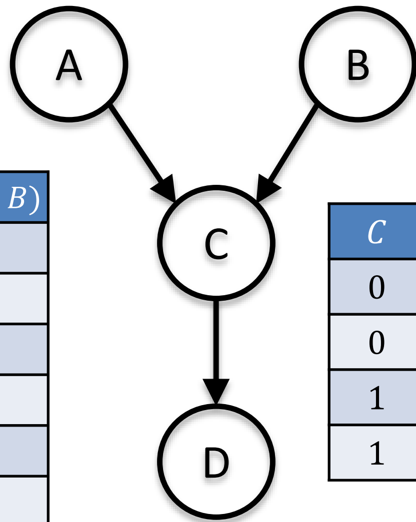
A	B	C	D
0	0	0	0
0	1	1	0

(1) Sample from $p(x_D|\ x_C = 1)$

$$p(x_D = 0|x_C = 1) = .4$$
$$p(x_D = 1|x_C = 1) = .6$$

Gibbs Sampling

A	$P(A)$
0	.3
1	.7



B	$P(B)$
0	.4
1	.6

Order: A, B, C, D, A, B, C, D, …

A	B	C	$P(C\|A,B)$
0	0	0	.1
0	0	1	.9
0	1	0	.2
0	1	1	.8
1	0	0	.01
1	0	1	.99
1	1	0	.25
1	1	1	.75

C	D	$P(D\|C)$
0	0	.3
0	1	.7
1	0	.4
1	1	.6

A	B	C	D
0	0	0	0
0	1	1	0

(2) Repeat the same process to generate the next sample

Gibbs Sampling

- Gibbs sampling forms a Markov chain
- The states of the chain are the assignments and the probability of transitioning from an assignment y to an assignment z (in the order $1, \ldots, n$)

$$p\left(z_1 \middle| y_{V\setminus\{1\}}\right) p\left(z_2 \middle| y_{V\setminus\{1,2\}}, z_1\right) \ldots p\left(z_n \middle| z_{V\setminus\{n\}}\right)$$

- If there are no zero probability states, then the chain is irreducible and aperiodic (hence it converges)
- The stationary distribution is $p(x)$
  - Proof?

Gibbs Sampling

- Recall that it takes time to reach the steady state distribution from an arbitrary starting distribution

- The **mixing time** is the number of samples that it takes before the approximate distribution is close to the steady state distribution

  - In practice, this can take 1000s of iterations (or more)

  - We typically ignore the samples for a set amount of time called the **burn in phase** and then begin producing samples

Gibbs Sampling

- We can use Gibbs sampling for MRFs as well!
  - We don't need to compute the partition function to use it (why not?)
  - Many "real" MRFs will have lots of zero probability assignments
    - If you don't start with a non-zero assignment, the algorithm can get stuck (changing a single variable may not allow you to escape)
    - Might not be possible to go between all possible non-zero assignments by only flipping one variable at a time

Metropolis-Hastings Algorithm

- This idea of choosing a transition probability between new assignments and the current assignments can be generalized beyond the transition probabilities used in Gibbs sampling

- Pick some transition function $q(x'|x)$ dependent on the current state x

  – Again, we would ideally want the probability of transitioning between any two non-zero states to be non-negative

Metropolis-Hastings Algorithm

- Choose an initial assignment x
- Sample an assignment z from the proposal distribution $q(x'|x)$
- Sample r uniformly from $[0,1]$
- If $r < \min\left\{1, \frac{p(z)q(x|z)}{p(x)q(z|x)}\right\}$
  - Set x to z
- Else
  - Leave x unchanged

Metropolis-Hastings Algorithm

- Choose an initial assignment x
- Sample an assignment z from the proposal distribution $q(x'|x)$
- Sample r uniformly from $[0,1]$
- If $r < \min\left\{1, \frac{p(z)q(x|z)}{p(x)q(z|x)}\right\}$
  - Set x to z
- Else
  - Leave x unchanged

$\frac{p(z)}{q(z|x)}$ and $\frac{p(x)}{q(x|z)}$ are like importance weights

The acceptance probability is then a function of the ratio of the importance of z and the importance of x

Metropolis-Hastings Algorithm

- The Metropolis-Hastings algorithm produces a Markov chain that converges to $p(x)$ from any initial distribution (assuming that it is irreducible and aperiodic)

- What are some choices for $q(x'|x)$?

  – Use an importance sampling distribution

  – Use a uniform distribution (like a random walk)

- Gibbs sampling is a special case of this algorithm where the acceptance probability is always equal to one