

CS 3750 Machine Learning Lecture 11

Monte Carlo inference

Milos Hauskrecht
milos@cs.pitt.edu
5329 Sennott Square

CS 3750 Advanced Machine Learning

Markov chain Monte Carlo

- Likelihood weighting: samples are generated according to Q and every sample from Q is reweighted according to its likelihood, but the Q distribution may be very far from the target
- MCMC is a strategy for generating samples from the target distribution, including conditional distributions
- MCMC:
 - Markov chain defines a sampling process that
 - initially generates samples very different from the target posterior
 - but gradually refines the samples so that they are closer and closer to the posterior.

CS 3750 Advanced Machine Learning

MCMC

- The construction of a Markov chain requires two basic ingredients
 - a transition matrix
 - an initial distribution
- Assume a finite set $S = \{1, \dots, m\}$ of states, then a transition matrix is

$$\mathbf{P} = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mm} \end{pmatrix}$$

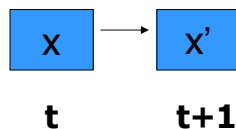
Where $p_{ij} \geq 0 \quad \forall (i, j) \in S^2$ and $\sum_{j \in S} p_{ij} = 1 \quad \forall i \in S$

Markov Chain

- Markov chain defines a random process:

$$x^{(0)}, x^{(1)}, \dots, x^{(m)}, \dots$$

Initial state



- Chain Dynamics

$$P^{(t+1)}(X^{(t+1)} = x') = \sum_{x \in \text{Dom}(X)} P^{(t)}(X^{(t)} = x) T(x \rightarrow x')$$

Probability of a state x' at time $t+1$

MCMC

- Markov chain satisfies

$$P[X_{n+1} = j | X_0 = i_0, \dots, X_n = i_n] = P[X_{n+1} = j | X_n = i_n]$$

- Irreducibility: A MC is called irreducible (or undecomposable) if there is a positive transition probability for all pairs of states within a limited number of steps
- In irreducible chains there may still exist a periodic structure such that for each state $i \in \mathcal{S}$, the set of possible return times to i when starting in i is a subset of the set $p \cdot \mathbf{N} = \{p, 2p, 3p, \dots\}$ containing all but a finite set of these elements. The smallest number p with this property is the so-called period of the chain

$$p = \text{gcd}\{n \in \mathbf{N} : p_{ii}^{(n)} > 0\}.$$

MCMC

- **Aperiodicity:** An irreducible chain is called aperiodic (or acyclic) if the period p equals 1 or, equivalently, if for all pairs of states there is an integer n_{ij} such that for all $n \geq n_{ij}$, the probability $p_{ij}^{(n)} > 0$.
- If a Markov chain satisfy both irreducibility and aperiodicity, then it **converges to an invariant distribution $q(x)$**
- A Markov chain with transition matrix P will have an equilibrium distribution q **iff** $q = qP$.
- A sufficient, but not necessary, condition to ensure a particular $q(x)$ is the invariant distribution of transition matrix P is the following reversibility (detailed balance) condition

$$q(x^i)P(x^{i-1} | x^i) = q(x^{i-1})P(x^i | x^{i-1})$$

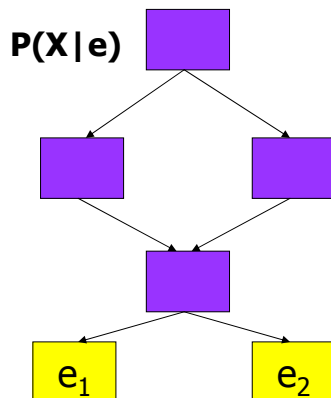
Markov Chain Monte Carlo

Objective: generates samples from the posterior distribution

• **Idea:**

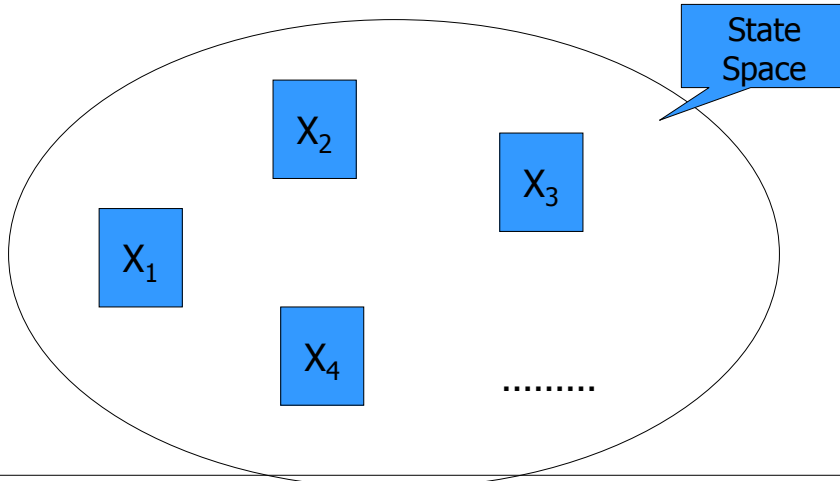
- Markov chain defines a sampling process that
- initially generates samples very different from the target posterior
- but gradually refines the samples so that they are closer and closer to the posterior.

MCMC



- $P(X|e)$ — the query we want to compute
- e_1 & e_2 are known evidence
- Sampling from the distribution $P(X)$ is very different from the desired posterior $P(X|e)$

Markov Chain Monte Carlo (MCMC)



CS 3750 Advanced Machine Learning

MCMC (Cont.)

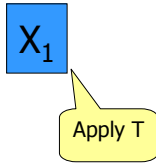
- **Goal:** a sample from $P(X|e)$
- Start from some $P(X)$ and generate a sample x_1

X_1

CS 3750 Advanced Machine Learning

MCMC (Cont.)

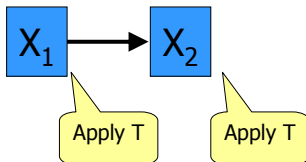
- **Goal:** a sample from $P(X|e)$
- Start from some $P(X)$ and generate a sample x_1



CS 3750 Advanced Machine Learning

MCMC (Cont.)

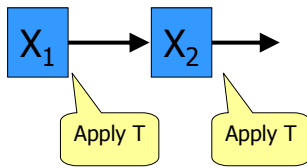
- **Goal:** a sample from $P(X|e)$
- Start from some $P(X)$ and generate a sample x_1



CS 3750 Advanced Machine Learning

MCMC (Cont.)

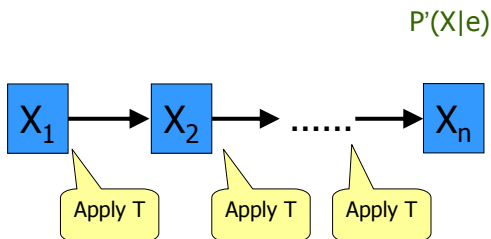
- **Goal:** a sample from $P(X|e)$
- Start from some $P(X)$ and generate a sample x_1



CS 3750 Advanced Machine Learning

MCMC (Cont.)

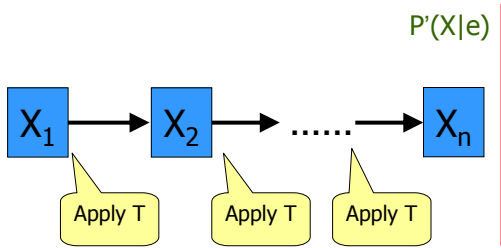
- **Goal:** a sample from $P(X|e)$
- Start from some $P(X)$ and generate a sample x_1



CS 3750 Advanced Machine Learning

MCMC (Cont.)

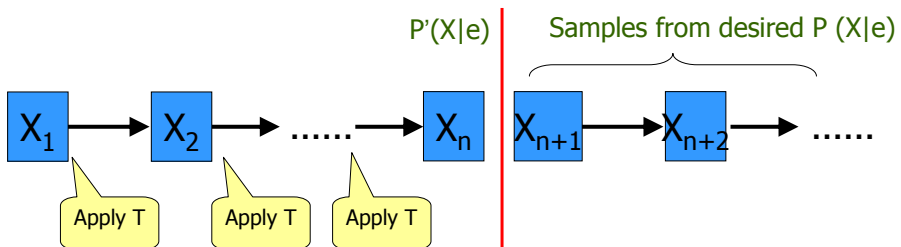
- **Goal:** a sample from $P(X|e)$
- Start from some $P(X)$ and generate a sample x_1



CS 3750 Advanced Machine Learning

MCMC (Cont.)

- **Goal:** a sample from $P(X|e)$
- Start from some $P(X)$ and generate a sample x_1



CS 3750 Advanced Machine Learning

MCMC

- MCMC sampling process doesn't converge to a stationary distribution definitely
 - Stationary distribution
- The stationary distribution is not unique, it depends on initial states

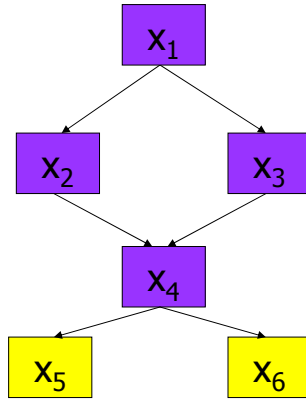
$$\pi(X = x') = \sum_{x \in \text{Dom}(X)} \pi(X = x) T(x \rightarrow x')$$

MCMC

- In general, an MCMC sampling process doesn't have to converge to a stationary distribution
- A finite state Markov Chain has a unique stationary distribution iff the markov chain is regular
 - regular: exist some k, for each pair of states x and x', the probability of getting from x to x' in exactly k steps is greater than 0
- We want Markov chains that converge to a unique target distribution from any initial state
- How to build such Markov chains?

Gibbs Sampling

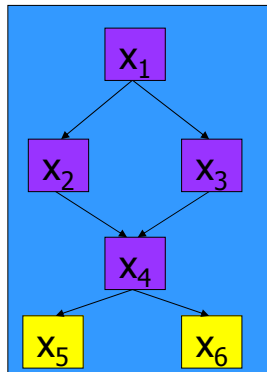
- a simple method to define MC for BBN can benefit from the structure (independences) in the network



- Evidence:
 - $x_5 = T$
 - $x_6 = T$
- all variables have binary values T or F

Gibbs Sampling

Initial state



\mathbf{x}_0

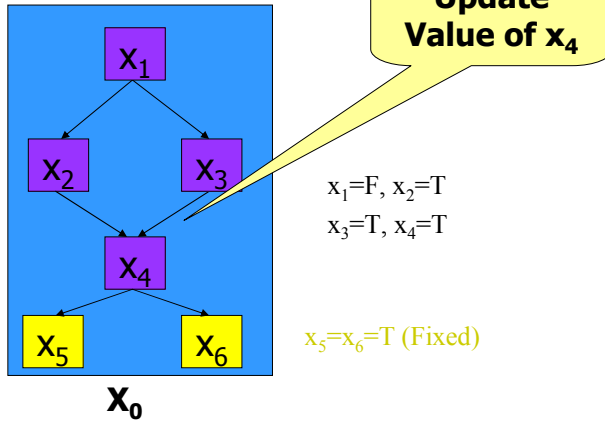
$x_1 = F, x_2 = T$

$x_3 = T, x_4 = T$

$x_5 = x_6 = T$ (Fixed)

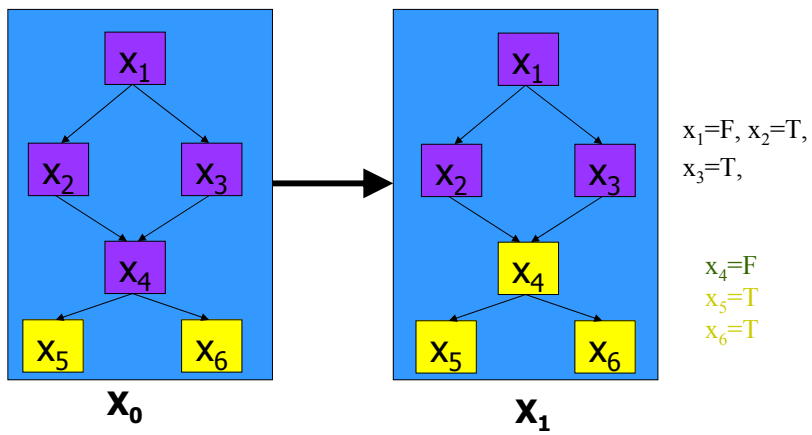
Gibbs Sampling

Initial state



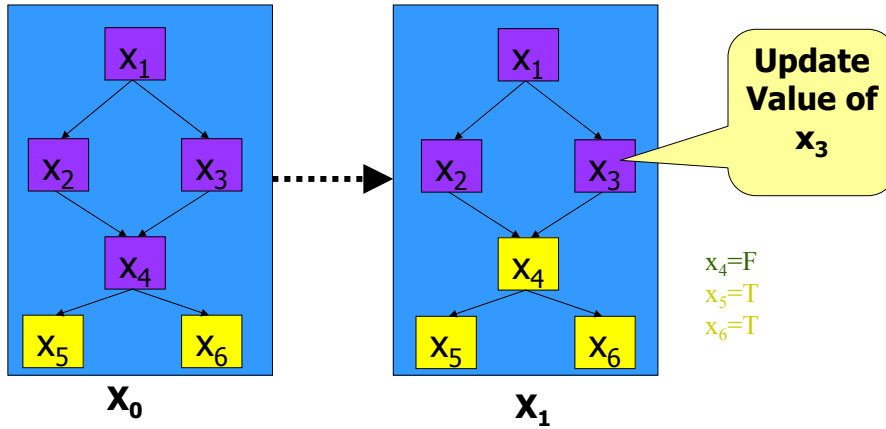
CS 3750 Advanced Machine Learning

Gibbs Sampling



CS 3750 Advanced Machine Learning

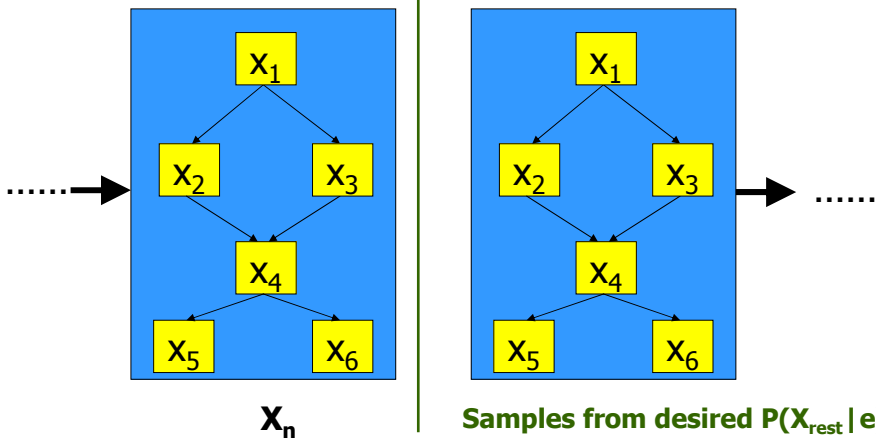
Gibbs Sampling



CS 3750 Advanced Machine Learning

Gibbs Sampling

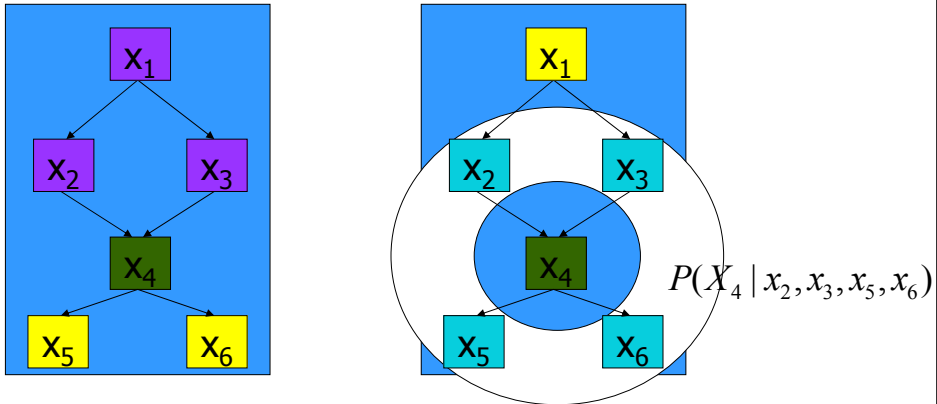
After many reassignments



CS 3750 Advanced Machine Learning

Gibbs Sampling

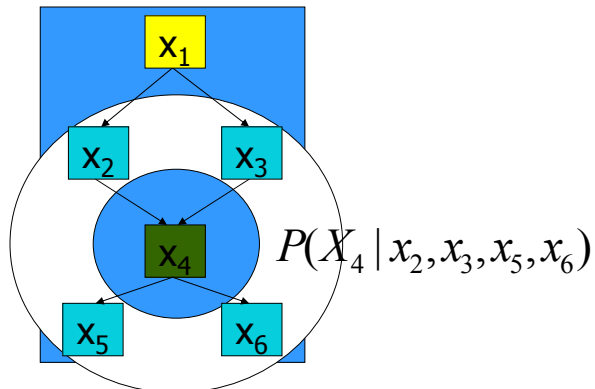
Keep resampling each variable using the value of variables in its local neighborhood (Markov blanket)



CS 3750 Advanced Machine Learning

Gibbs Sampling

- Gibbs sampling takes advantage of the structure
- Markov blanket makes the variable independent from the rest of the network



CS 3750 Advanced Machine Learning

Building a Markov Chain

- **A reversible Markov chain:**
- A sufficient, but not necessary, condition to ensure a particular $q(x)$ is the invariant distribution of transition matrix P is the following reversibility (detailed balance) condition

$$q(x^i)P(x^{i-1} | x^i) = q(x^{i-1})P(x^i | x^{i-1})$$

- **Metropolis-Hastings algorithm**
 - builds a reversible Markov Chain
 - Uses a proposal distribution to generate candidate states
 - Either accept it and take a transition to state x'
 - Or reject it and stay at current state x

Building a Markov Chain

- **Metropolis-Hastings algorithm**
 - builds a reversible Markov Chain
 - uses the **proposal distribution** (similar to proposal the distribution in importance sampling) to generate candidates for x'
 - A proposal distribution $Q: T^Q(x \rightarrow x')$
 - Example: Uniform over the values of variables
 - Either accept proposal and take a transition to state x'
 - Or reject it and stay at current state x
 - Acceptance probability

$$A(x \rightarrow x')$$

Building a Markov Chain

- Transition for the MH:

$$T(x \rightarrow x') = T^Q(x \rightarrow x') A(x \rightarrow x') \quad \text{if } x \neq x'$$

$$T(x \rightarrow x) = T^Q(x \rightarrow x) + \sum_{x' \neq x} T^Q(x \rightarrow x')(1 - A(x \rightarrow x'))$$

otherwise

- From reversibility condition:

$$\pi(x)T(x \rightarrow x') = \pi(x')T(x' \rightarrow x)$$

- We get

$$A(x \rightarrow x') = \min\left[1, \frac{\pi(x')T^Q(x' \rightarrow x)}{\pi(x)T^Q(x \rightarrow x')}\right]$$

Building a Markov Chain

- Comparing MH with Gibbs

- For Gibbs

$$A(u_i, x_i \rightarrow u_i, x'_i)$$

$$= \min\left[1, \frac{P(x'_i | u_i)T^Q(u_i, x'_i \rightarrow u_i, x_i)}{P(x_i | u_i)T^Q(u_i, x_i \rightarrow u_i, x'_i)}\right]$$

$$= \min\left[1, \frac{P(x'_i | u_i)P(x_i | u_i)}{P(x_i | u_i)P(x'_i | u_i)}\right]$$

$$= \min[1, 1] = 1$$

- Special MH, for which acceptance probability is 1.

MH algorithm

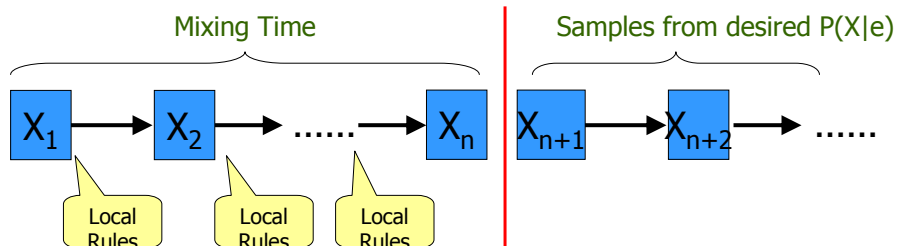
- **Assumptions:**
 - We can't draw samples from $q(x)$
 - We can evaluate $q(x)$ for any x
- We use a Markov chain that moves towards x^* with acceptance probability

$$A(x, x^*) = \min \left[1, \frac{q(x^*)p(x | x^*)}{q(x)p(x^* | x)} \right]$$

- The transition kernel defined by this process satisfies the detailed balance condition

Mixing Time in Using Markov Chain

- Mixing Time
 - The number of steps we take until we collect a sample from the target distribution. ($\# = n$)



Summary

- Markov Chain Monte Carlo method attempts to generate samples from posterior distribution
- Metropolis Hastings algorithm is a general scheme for specifying a Markov chain.
- Gibbs sampling is a special case that takes advantage of the network structure (Markov Blanket)