## Lecture 14: Particle-based inference: Gibbs sampling

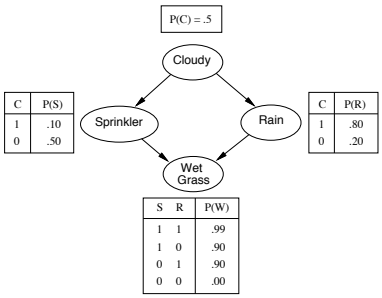
- Gibbs sampling
- Markov chains
- Markov Chain Monte Carlo (MCMC) methods



## Recall: Particle-based inference

- Suppose we have evidence $E = e$ and we want to know $p(Y|E = e)$ for some query variables $Y$
- Particle-based methods will generate particles and then compute sufficient statistics to estimate this answer
- Likelihood weighting has an easy way of producing samples: go through the Bayes net in the direction of the arcs, sample nodes without evidence and set the value for evidence variables
- Since these samples are NOT from $p(Y|E = e)$ each particle must have a weight. The weights are used instead of counts in the probability estimation.
- But these weights can get very small, and then we would need to sample a lot of data to get good estimates.

## A different idea




- Suppose we want to compute $P(R|S=1)$
- We generate *one sample*, with the given evidence variables instantiated correctly
- *Then we keep changing it!*
- If we are careful, we will get samples from the correct distribution



## Gibbs sampling

1. Initialization
   - Set evidence variables $E$, to the observed values $e$
   - Set all other variables to random values (e.g. by forward sampling, uniform sampling...)

   This gives us a sample $x_1, \ldots, x_n$.
2. Repeat (as much as wanted)
   - Pick a non-evidence variable $X_i$ uniformly randomly)
   - Sample $x_i'$ from $p(X_i|x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$.
   - Keep all other values: $x_j' = x_j, \forall j \neq i$
   - The new sample is $x_1', \ldots, x_n'$
3. Alternatively, you can march through the variables in some predefined order

## Why Gibbs works in Bayes nets

- The key step is sampling according to $p(X_i|x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$. How do we compute this?
- In Bayes nets, we know that a variable is conditionally independent of all others given its Markov blanket (parents, children, spouses)

$$p(X_i|x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = p(X_i|\text{MarkovBlanket}(X_i))$$

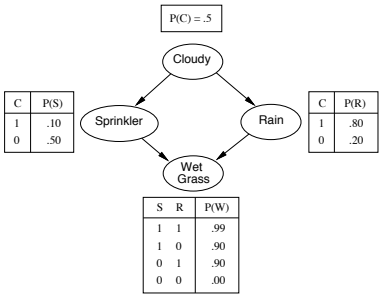
- So we need to sample from $p(X_i|\text{MarkovBlanket}(X_i))$
- Let $Y_j, j = 1, \ldots, k$ be the children of $X_i$. It is easy to show



that:

$$p(X_i = x_i|\text{MarkovBlanket}(X_i)) \propto p(X_i = x_i|\text{Parents}(X_i)) \cdot \cdot \prod_{j=1}^{k} p(Y_j = y_j|\text{Parents}(Y_j))$$

## Example




1. Generate a first sample: $C = 0, R = 0, S = 0, W = 1$.
2. Pick $R$, sample it from $p(R|C = 0, W = 1, S = 0)$. Suppose we get $R = 1$.
3. Our new sample is $C = 0, R = 1, S = 0, W = 1$
4. ....



## Analyzing Gibbs sampling

- Consider the variables $X_1, \ldots, X_n$. Each possible assignment of values to these variables is a state of the world, $\langle x_1, \ldots, x_n \rangle$.
- In Gibbs sampling, we start from a given state $s = \langle x_1, \ldots, x_n \rangle$. Based on this, we generate a new state, $s' = \langle x'_1, \ldots, x'_n \rangle$.
- *$s'$ depends only on $s$*!
- There is a well-defined probability of going from $s$ to $s'$.

Gibbs sampling constructs a **Markov chain** over the Bayes net

## Markov chains

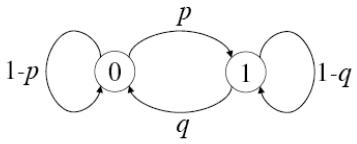
- Suppose you have a system which evolves through time:

$$s_0 \rightarrow s_1 \rightarrow \cdots \rightarrow s_t \rightarrow s_{t+1} \rightarrow \ldots$$

- A **Markov chain** is a special case of such a system, defined by:
  - A set of states $S$
  - A starting distribution over the set of states $p_0(s) = p(s_0 = s)$. If the state space is discrete, this can be represented as a column vector $\mathbf{p}_0$
  - A stationary transition probability, specifying $\forall s, s' \in S$, $p_{ss'} = p(s_{t+1} = s' | s_t = s)$. The **Markov property** here means that $p(s_{t+1}|s_t) = p(s_{t+1}|s_0, \ldots s_t)$.
- For convenience, we put these probabilities in a $|S| \times |S|$ **transition matrix** $\mathbf{T}$.



## Example of a Markov chain




- State space $S = \{0, 1\}$
- Transition matrix:

$$\mathbf{T} = \begin{bmatrix} (1-p) & p \\ q & (1-q) \end{bmatrix}$$

- We can fix an initial probability distribution, e.g. $\mathbf{p}_0 = \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix}$

## Example: Single-server queue

- Consider a single checkout at the grocery store
  - A single customer arrives at any given times step with probability $p$
  - The customer at the head of the line is served with probability $q$ on any given time step
  - Multiple customers can arrive at the same time
- The state of the chain is given by the number of customers in the queue
- There is a well-defined transition probability from any given state to any other state, which can be computed from $p$ and $q$ (what is it?)



## How does the chain evolve over time?

- Where will the chain be on the first time step, $t = 1$?

$$p(s_{t+1} = s') = \sum_s p(s_0 = s)p(s_1 = s'|s_0 = s)$$

  by using the graphical model for the first time step: $s_0 \longrightarrow s_1$.
- We can put this in matrix form as follows:

$$\mathbf{p}'_1 = \mathbf{p}'_0\mathbf{T} \longrightarrow \mathbf{p}_1 = \mathbf{T}'\mathbf{p}_0$$

  where $\mathbf{T}'$ denotes the transpose of $\mathbf{T}$
- Similarly, at $t = 2$, we have:

$$\mathbf{p}_2 = \mathbf{T}'\mathbf{p}_1 = (\mathbf{T}')^2\mathbf{p}_0$$

## Steady-state (stationary) distribution

- By induction, the probability distribution over possible states at time step $t$ can be computed as:

$$\mathbf{p}_t = \mathbf{T}'\mathbf{p}_{t-1} = (\mathbf{T}')^t\mathbf{p}_0$$

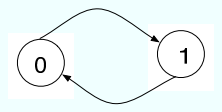
- If $\lim_{t\to\infty} \mathbf{p}_t$ exists, it is called the **stationary or steady-state distribution** of the chain.
- If the limit exists, $\pi = \lim_{t\to\infty} \mathbf{p}_t$, then we have:

$$\pi = \mathbf{T}'\pi, \sum_{s\in S} \pi_s = 1$$

- Under what conditions does a chain have a stationary distribution?
- Does the equation $\pi = \mathbf{T}'\pi$ always have a unique solution?

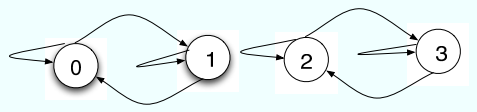


## Not all chains have a stationary distribution




- If the chain has a purely periodic cycle, the stationary distribution does not exist
- E.g. in the chain above, the system is always in one state on odd time steps and the other state on even time steps, so the probability vector $\mathbf{p}_t$ oscillates between 2 values
- For the limit to exist, the chain must be **aperiodic**
- A standard trick for breaking periodicity is to add self-loops with small probability

## Limit distribution may depend on the initial transition




- If the chain has multiple "components", the limit distribution may exist, but depend on a few initial steps
- E.g. if all transitions above have probability 0.5, there are two possible stationary distributions: $[0.5\ 0.5\ 0\ 0]$ and $[0\ 0\ 0.5\ 0.5]$
- Such a chain is called **reducible**
- To eliminate this, every state must be able to reach every other state:

$$\forall s, s', \exists k > 0 \text{ s.t. } p(s_{t+k} = s' | s_t = s) > 0$$



## Ergodicity

- An **ergodic** Markov chain is one in which any state is reachable from any other state, and there are no strictly periodic cycles (in other words, the chain is irreducible and aperiodic)
- In such a chain, there is a unique stationary distribution $\pi$, which can be obtained as:

$$\pi = \lim_{t \to \infty} \mathbf{p}_t$$

  This is also called the **equilibrium** distribution
- The chain reaches the equilibrium distribution regardless of $\mathbf{p}_0$
- The distribution can be computed by solving:

$$\pi = \mathbf{T}'\pi, \sum_s \pi_s = 1$$

## Balance in Markov chains

- Consider the steady-state equation for a system of $n$ states:

$$[\pi_1 \pi_2 \ldots \pi_n] = [\pi_1 \pi_2 \ldots \pi_n] \begin{bmatrix} 1 - \sum_{i \neq 1} p_{1i} & p_{12} & \ldots & p_{1n} \\ p_{21} & 1 - \sum_{i \neq 2} p_{2i} & \ldots & p_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ p_{n1} & p_{n2} & \ldots & 1 - \sum_{i \neq n} p_{ni} \end{bmatrix}$$

- By doing the multiplication, for any state $s$, we get:

$$\pi_s = \pi_s \left(1 - \sum_{i \neq s} p_{si}\right) + \sum_{i \neq s} \pi_i p_{is} \Longrightarrow \pi_s \sum_{i \neq s} p_{si} \sum_{i \neq s} \pi_i p_{is}$$

  This can be viewed as a "flow" property: the flow out of $s$ has to be equal to the flow coming into $s$ from all other states



## Detailed balance

- Suppose we were designing a Markov chain, and we wanted to ensure a stationary distribution
- This means that the flow equilibrium at every state must be achieved.
- One way to ensure this is to make flow equal between *any pair* of states:

$$\pi_s p_{ss'} = \pi_{s'} p_{s's}$$

  This gives us a *sufficient condition* for stationarity, called **detailed balance**
- A Markov chain with this property is called **reversible**

## Mixing time

- Sometimes, instead of computing the stationary distribution of a chain, we would like to *sample* from it
- Hence, it is useful to know for what value of $t$ we have $\mathbf{p}_t \approx \pi$
- This is called the **mixing time** of the chain
- There are different ways to measure the approximate difference of these two distributions
- If the mixing time is "small" (e.g. compared to the number of states in the chain) we say the chain is **rapidly mixing**



## Markov Chain Monte Carlo (MCMC) methods

- Suppose you want to generate samples from some distribution, but it is hard to get samples directly
  E.g., We want to sample uniformly the space of graphs with certain properties
- You set up a Markov chain such that its *stationary distribution* is the *desired distribution*
- Note that the 'states" of this chain can be fairly complicated!
- You start at some state, let time pass, and then take samples
- For this to work we need to ensure that:
  - the chain has a unique stationary distribution
  - the stationary distribution is what we want
  - we reach the stationary distribution quickly

## Sampling the equilibrium distribution

- We can sample $\pi$ just by running the chain a long time:
  - Set $s_0 = i$ for some arbitrary $i$
  - For $t = 1, \ldots, M$, if $s_t = s$, sample a value $s'$ for $s_{t+1}$ based on $p_{ss'}$
  - Return $s_M$.

  If $M$ is large enough, this will be a sample from $\pi$
- In practice, we would like to have a rapidly mixing chain, i.e. one that reaches the equilibrium quickly



## Example: Random graphs

- Suppose you want to sample uniformly from the space of graphs with $v$ vertices and certain properties (e.g. certain in-degree and out-degree bounds, cycle properties...)
- You set up a chain whose states are graphs with $v$ vertices
- Transitions consist of adding or removing an arc (reversal too, if the graphs are directed), with a certain probability
- We start with a graph satisfying the desired property.
- The probabilities are devised based on the distribution that we want to reach in the limit.

## MCMC for sampling from a graphical model

- The states of the chain are instances, in which the evidence variables are instantiated to their known values
- Transitions allow changing the value of a non-evidence variable in the instance
- The stationary distribution has to be the conditional distribution of the model given the evidence
- This is ensured by specifying the transition matrix of the chain based on the original model.
- Gibbs sampling is an example of this approach



## Implementation issues

- The initial samples are influenced by the starting distribution, so they need to be thrown away. This is called the **burn-in stage**
- Because burn-in can take a while, we would like to draw several samples from the same chain
- However, if we take samples $t$, $t+1$, $t+2$..., they will be highly correlated
- Usually we wait for burn-in, then take every $n$th sample, for some $n$ sufficiently large. This will ensure that the samples are (for all practical purposes) uncorrelated

## Gibbs sampling as MCMC

- We have a set of random variables $X = \{x_1 \ldots x_n\}$, with evidence variables $E = e$. We want to sample from $p(X - E|E = e)$.
- Let $X_i$ be the variable to be sampled, currently set to $x_i$, and $\bar{x}_i$ be the values for all other variables in $X - E - \{X_i\}$
- The transition probability for the chain is: $p_{ss'} = p(x'_i|\bar{x}_i, e)$
- Under mild assumptions on the original graphical model, the chain is ergodic
- We want to show that $p(X - E|e)$ is the stationary distribution



## Gibbs satisfies detailed balance

- We show that if we plug in $p(X - E|e)$ as the stationary distribution, detailed balance is satisfied

$$\begin{aligned}
\pi_s p_{ss'} &= p(X - E|e)p(x'_i|\bar{x}_i, e) \\
&= p(x_i, \bar{x}_i|e)p(x'_i|\bar{x}_i, e) \\
&= p(x_i|\bar{x}_i, e)p(\bar{x}_i|e)p(x'_i|\bar{x}_i, e) \text{ (by chain rule)} \\
&= p(x_i|\bar{x}_i, e)p(x'_i, \bar{x}_i|e) \text{ (backwards chain rule)} \\
&= p_{s's}\pi_{s'}
\end{aligned}$$

- If the chain is ergodic, there is a unique stationary distribution, and since $p(X - E|e)$ satisfies the balance equation, it must be the stationary distribution.