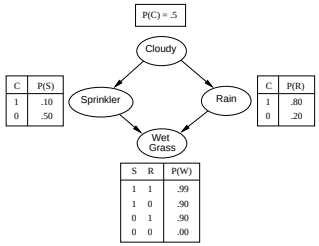
Lecture 12: Particle-based inference: Gibbs sampling

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

Recall from last time: 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 from a "mutilated" Bayes net, 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 Ej, 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

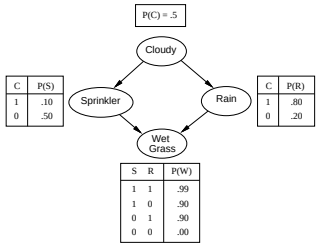
- The key step is sampling according to $p(X_i|x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n)$. How do we compute this?
- In Bayes nets, we know that a variable is conditionally independent of all other *given its Markov blanket* (parents, children, spouses)

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

- 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. We can 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

A Markov chain is 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 $p_{ss'} = p(s_{t+1} = s'|s_t = s)$. For convenience, we often put these in a $n \times n$ matrix T

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

Steady-state (stationary) distribution

- Where will the chain be in 1 step?

$$\mathbf{p_1'} = \mathbf{p_0'}T \longrightarrow \mathbf{p_1} = T'\mathbf{p_0}$$

  where T' denotes the transpose of T
- In two steps?

$$\mathbf{p_2} = T'\mathbf{p_1} = (T')^2\mathbf{p_0}$$

- In t steps?

$$\mathbf{p_t} = T'\mathbf{p_{t-1}} = (T')^t\mathbf{p_0}$$

  A **stationary distribution** π is a distribution left invariant by the chain:$\pi = T'\pi$

Properties of Markov chains

- Do all Markov chains converge to a stationary distribution?
- When this distribution exists, is it always unique?
- Are there conditions under which we can guarantee that the distribution is unique?

Properties of Markov chains

- Do all Markov chains converge to a stationary distribution?
  No! Consider **periodic** Markov chains (which contain cycles)
- When this distribution exists, is it always unique?
  No! It may depend on the initial distribution. Such chains are called **reducible**
- Are there conditions under which we can guarantee that the distribution is unique?
  Yes.

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 such a chain, there is a unique stationary distribution π, which can be obtained as:

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

  This is called **equilibrium** distribution
- Note that the chain reaches the equilibrium distribution *regardless of* $\mathbf{p}_0$

Detailed balance

- Consider the stationary distribution:

$$\pi(s') = \sum_s \pi(s)p(s,s')$$

  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 states
- One way to ensure this is to make flow equal between *any pair* of states:

$$\pi(s)p(s,s') = \pi(s')p(s',s)$$

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

Markov Chain Monte Carlo 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
  - that the stationary distribution is what we want
  - that we reach the stationary distribution quickly

Sampling the equilibrium distribution

- We can sample π 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 $p(s,s')$
  - Return s_M.

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

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 nth 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: $p(s, s') = p(x'_i|\bar{x}_i, e)$
- Obviously the chain is ergodic
- We want to show that $p(X - E|e)$ is the stationary distribution.

Gibbs satisfies detailed balance

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