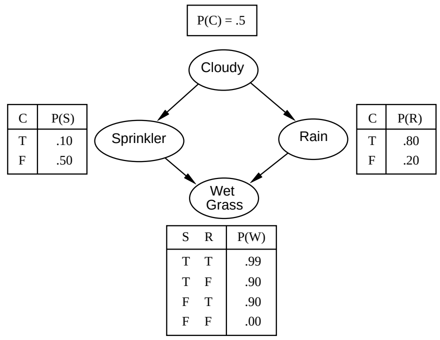
Lecture 13: Markov Chain Monte Carlo. Gibbs sampling

- Gibbs sampling
- Markov chains

Recall: Approximate inference using samples

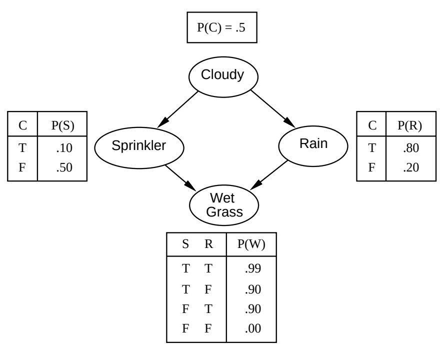
- Main idea: we generate samples from our Bayes net, then compute probabilities using (weighted) counts)
- But we may need a lot of work to get enough samples (e.g. if the CPDs are very extreme)
- Rejection sampling and likelihood weighting are also specific to directed models

Recall: Forward sampling



- Each sample is constructed from scratch!
- We saw two alternatives for incorporating the evidence
  - Throw away samples that are inconsistent with it
  - Force the evidence variables, but then weigh the samples
- In both cases, after a sample is constructed, we start a new one from scratch

A different idea



- Suppose we want to compute $P(R|S=t)$
- 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 Z_j, to the observed values z_j
   - 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

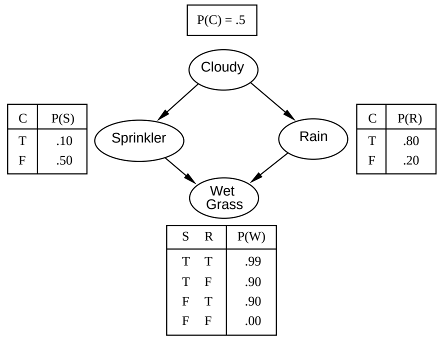
- The key step is sampling according to $P(X_i|x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$. How do we compute this?
- In Bayes nets, we know that a variable is conditionally independent of all other *given its Markov blanket* (parents, children, spouses)

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

- 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. You will show (next homework) that:

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

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 $X_1, \dots, X_n$. Each possible assignment of values to these variables is a state of the world, $\langle x_1, \dots, x_n \rangle$.
- In Gibbs sampling, we start from a given state $s = \langle x_1, \dots, x_n \rangle$. Based on this, we generate a new state, $s' = \langle x'_1, \dots, x'_n \rangle$.
- *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 $p_0(s) = P(s_0 = s)$. We often put these in a 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 a $n \times n$ matrix P

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

Steady-state (stationary) distribution

- Where will the chain be in 1 step?

$$p_1^T = p_0^T P \longrightarrow p_1 = P p_0$$

- In two steps?

$$p_2 = P p_1 = P^2 p_0$$

- In t steps?

$$p_t = P p_{t-1} = P^t p_0$$

A **stationary distribution** π is a distribution left invariant by the chain:

$$\pi = P\pi$$

- Note that some chains can have more than one stationary distribution!

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** (why)

- A Markov chain with this property is called **reversible**

Monte Carlo Markov Chain (MCMC)

- Suppose we want to sample data from some distribution
- We will *set up a Markov chain* which has the desired distribution as its stationary distribution!
- For this we would like the chain to have a *unique* stationary distribution, so that we can get samples from it *regardless of the starting distribution*

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:

$$\pi = \lim_{t \to \infty} p_t$$

  This is called **equilibrium** distribution
- Note that the chain reaches the equilibrium distribution *regardless of p_0*

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, you'd 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-it can take a while, we'd 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 $\mathbf{X} = \{x_1 \ldots x_n\}$, with evidence variables $\mathbf{Z} = \mathbf{z}$. We want to sample from $p(\mathbf{X} - \mathbf{Z}|\mathbf{z})$.
- Let X_i be the variable to be sampled, currently set to x_i, and $\bar{\mathbf{x}}_\mathbf{i}$ be the values for all other variables in $X - Z - \{X_i\}$
- The transition probability for the chain is: $P(s, s') = p(x_i'|\bar{\mathbf{x}}_\mathbf{i}, \mathbf{z})$
- Obviously the chain is ergodic
- We want to show that $p(\mathbf{X} - \mathbf{Z}|\mathbf{z})$ is the stationary distribution.

Gibbs satisfies detailed balance

$$\begin{aligned}
\pi(s)P(s,s') &= p(\mathbf{X}-\mathbf{Z}|\mathbf{z})p(x_i'|\bar{\mathbf{x}}_{\mathbf{i}},\mathbf{z}) \\
&= p(x_i,\bar{\mathbf{x}}_{\mathbf{i}}|\mathbf{z})p(x_i|\bar{\mathbf{x}}_{\mathbf{i}},\mathbf{z}) \\
&= p(x_i|\bar{\mathbf{x}}_{\mathbf{i}},\mathbf{z})p(\bar{\mathbf{x}}_{\mathbf{i}}|\mathbf{z})p(x_i'|\bar{\mathbf{x}}_{\mathbf{i}},\mathbf{z}) \text{ (by chain rule)} \\
&= p(x_i|\bar{\mathbf{x}}_{\mathbf{i}},\mathbf{z})p(x_i',\bar{\mathbf{x}}_{\mathbf{i}}|\mathbf{z}) \text{ (backwards chain rule)} \\
&= P(s',s)\pi(s)
\end{aligned}$$