Monte Carlo Markov Chain

MCMC

Limitations of LW:

- Evidence affects sampling only for nodes that are its descendants
- For nondescendants, the weights account for the effect of the evidence
- If evidence is at the leaves, we are sampling from the prior distribution (and not the posterior which is what we want)
MCMC

Strategy used by MCMC
• Generate a sequence of samples
• Initial samples generated from the prior
• Successive samples generated progressively closer to the posterior

Applies to both directed and undirected models. We’ll use a distribution $P_\Phi$ defined in terms of a set of factors $\Phi$.

Gibbs Sampling
Gibbs Sampling

Example: Suppose we have as evidence SAT = High and Letter = Weak (nodes are shaded grey)

Factors:
• \( P(I) \)
• \( P(D) \)
• \( P(G \mid I, D) \)

Reduced Factors:
• \( P(S=\text{high} \mid I) \)
• \( P(L=\text{weak} \mid G) \)

Eliminate all rows that are inconsistent with the evidence in all factors (see pg 111 of textbook)

Gibbs Sampling

Start with an initial sample eg: \( \mathbf{x}^{(0)} = (D = \text{high}, I = \text{low}, G = B, S = \text{high}, L = \text{weak}) \)

• \( D, I \) and \( G \) could be set in any way, for instance by forward sampling, to get \( D^{(0)} = \text{high}, I^{(0)} = \text{low}, G^{(0)} = B \)

• \( S=\text{high} \) and \( L=\text{weak} \) are observed
Gibbs Sampling

Resample non-evidence nodes, one at a time, in some order eg. G, I, D.

If we sample $X_i$, keep other nodes clamped at the values of the current state ($D = \text{high}, I = \text{low}, G = B, S = \text{high}, L = \text{weak}$)

To sample $G^{(1)}$, we compute $P_{\phi}(G \mid D=\text{high}, I=\text{low}, S=\text{high}, L=\text{weak})$:

$$
P_{\phi}(G \mid D=\text{high}, I=\text{low}, S=\text{high}, L=\text{weak})
= \frac{P(I = \text{high})P(D = \text{low})P(G \mid I = \text{low}, D = \text{high})P(L = \text{low} | G)P(S = \text{high} | I = \text{low})}{\sum_{G} P(G \mid D = \text{low}, D = \text{high})P(L = \text{low} | G)P(S = \text{high} | I = \text{low})}
= \frac{\sum_{G} P(G \mid I = \text{high}, D = \text{high})P(L = \text{low} | G)P(S = \text{high} | I = \text{low})}{\sum_{G} P(G \mid I = \text{high}, D = \text{high})P(L = \text{low} | G)P(S = \text{high} | I = \text{low})}
$$

Suppose we obtain $G^{(1)} = C$.

Now sample $I^{(1)}$ from $P_{\phi}(I \mid D=\text{high}, G=C, S=\text{high}, L=\text{weak})$. Note it is conditioned on $G^{(1)}=C$

Say we get $I^{(1)}=\text{high}$

Now sample $D^{(1)}$ from $P_{\phi}(D \mid G=C, I = \text{high}, S=\text{high}, L=\text{weak})$. Say you get $D^{(1)}=\text{high}$

The first iteration of sampling produces $x^{(1)} = (I = \text{high}, D = \text{high}, G = C, S=\text{high}, L=\text{weak})$

Iterate...
Gibbs Sampling

- \( P_\Phi(G \mid D=\text{high}, I=\text{low}, S=\text{high}, L=\text{weak}) \) takes downstream evidence \( L=\text{weak} \) into account (makes it closer to the posterior distribution \( P(X \mid e) \))
- Early on, \( P_\Phi(G \mid D=\text{high}, I=\text{low}, S=\text{high}, L=\text{weak}) \) very much like the prior \( P(X) \) because it uses values for \( I \) and \( D \) sampled from \( P(X) \)
- On next iteration, resampling \( I \) and \( D \) conditioned on new value of \( G \) brings the sampling distribution closer to the posterior
- Sampling distribution gets progressively closer and closer to the posterior

---

**Gibbs Sampling Procedure**

```plaintext
Procedure Gibbs-Sample ( 
    X // Set of variables to be sampled 
    \Phi // Set of factors defining \( P_\Phi \) 
    P^{(0)}(X), // Initial state distribution 
    T // Number of time steps 
)
1. Sample \( x^{(0)} \) from \( P^{(0)}(X) \)
2. for \( t=1, ..., T \)
3. \( x^{(t)} \leftarrow x^{(t-1)} \)
4. for each \( X_i \in X \)
5. \( \text{Sample } x_i^{(t)} \text{ from } P_\Phi(X_i \mid x_{-i}) \)
6. // Change \( X_i \) in \( x^{(t)} \)
7. return \( x^{(0)}, ..., x^{(T)} \)
```
Gibbs Sampling

Gibbs sampling with evidence
- Reduce all factors by the observations $e$
- The distribution $P_{\Phi}$ corresponds to $P(X|e)$

Markov Chains
Markov Chains

• (Informally) A Markov chain is a graph of states over which the sampling algorithm takes a random walk

• Note: the graph is not the graphical model but a graph over the possible assignments to a set of variables $X$

Markov Chains

• A Markov chain is defined via a state space $\text{Val}(X)$ and a model that defines, for every state $x \in \text{Val}(X)$ a next-state distribution over $\text{Val}(X)$.

• More precisely, the transition model $T$ specifies for each pair of states $x, x'$ the probability $T(x \rightarrow x')$ of going from $x$ to $x'$.

• A homogeneous Markov chain is one where the system dynamics do not change over time
Markov Chains

Example of a Markov Chain with Val(X)={A,B,C}:

State Transition Diagram View

Conditional Probability Distribution View

| $X_{t+1}$ | $X_t$ | $P(X_t|X_{t+1})$ |
|-----------|-------|-----------------|
| A         | A     | 0.25            |
| A         | B     | 0               |
| A         | C     | 0.75            |
| B         | A     | 0.5             |
| B         | B     | 0.5             |
| B         | C     | 0               |
| C         | A     | 0.4             |
| C         | B     | 0.6             |
| C         | C     | 0               |

Markov Chains

- Random sampling process defines a random sequence of states $x^{(0)}$, $x^{(1)}$, $x^{(2)}$, ...
- $X^{(t)}$ is a random variable:
- Need initial state distribution $P^{(0)}(X^{(0)})$
- Probability that next state is $x'$ can be computed as:

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

Sum over all states that the chain could have been at time $t$

Probability of transition from $x$ to $x'$


Markov Chains

How to generate a Markov Change Monte Carlo trajectory:

```plaintext
Procedure MCMC-Sample (P(0)(X), T, T)
    1. Sample x(0) from P(0)(X)
    2. for t = 1, ..., T
    3. Sample x(t) from T(x(t-1) → X)
    4. return x(0), ..., x(T)
```

The big question: does P(t) converge and what to?

• When the process converges, we expect:
  \[ P^{(t)}(x') \approx P^{(t+1)}(x') = \sum_{x \in Val(X)} P^{(t)}(x)T(x \to x') \]

• A distribution \( \pi(X) \) is a stationary distribution for a Markov chain \( T \) if it satisfies:
  \[ \pi(X = x') = \sum_{x \in Val(X)} \pi(X = x)T(x \to x') \]

• A stationary distribution is also called an invariant distribution
Markov Chains

Another example:

To find the stationary distribution:
\[ \pi(x_1) = 0.25\pi(x_1) + 0.5\pi(x_3) \]
\[ \pi(x_2) = 0.7\pi(x_2) + 0.5\pi(x_3) \]
\[ \pi(x_3) = 0.75\pi(x_1) + 0.3\pi(x_2) \]
\[ \pi(x_1) + \pi(x_2) + \pi(x_3) = 1 \]

Solving these simultaneous equations gives: \( \pi(x_1) = 0.2, \pi(x_2) = 0.5, \pi(x_3) = 0.3 \)

Markov Chains

- Bad news: no guarantee that MCMC sampling process converges to a stationary distribution
- Example of a periodic Markov chain (periodic = fixed cyclic behavior)
  - Start with \( P^{(0)}(x_1) = 1 \)
  - \( P^{(t)}(x_1) = 1 \) if \( t \) is even
  - \( P^{(t)}(x_2) = 1 \) if \( t \) is odd
Markov Chains

• No guarantee that stationary distribution is unique – depends on $P^{(0)}$
  – This happens if the chain is reducible: has states that are not reachable from each other

• We will restrict our attention to Markov chains that have a stationary distribution which is reached from any starting distribution $P^{(0)}$

To meet this restriction, we need the chain to be regular

• A Markov chain is said to be regular if there exists some number $k$ such that, for every $x, x' \in Val(X)$, the probability of getting from $x$ to $x'$ in exactly $k$ steps is > 0

• Theorem 12.3: If a finite state Markov chain $\mathcal{T}$ is regular, then it has a unique stationary distribution
Markov Chains

- Define $\mathcal{T}_i$ to be a transition model called a kernel
- For graphical models, define a kernel $\mathcal{T}_i$ for each variable $X_i \in \mathcal{X}$
- Define $\mathcal{X}_{-i} = \mathcal{X} - \{X_i\}$ and let $\mathbf{x}_i$ denote an instantiation to $X_i$
- The model $\mathcal{T}_i$ takes a state $(\mathbf{x}_{-i}, x_i)$ and transitions to a state $(\mathbf{x}_{-i}, x'_i)$

Gibbs Sampling Revisited
Gibbs Sampling Revisited

How do we use MCMC on a graphical model?

- Want to generate samples from the posterior
  \( P(\mathbf{x}|\mathbf{E}=\mathbf{e}) \) where \( \mathbf{X}=\mathbf{X}-\mathbf{E} \)
- Define a chain where \( P(\mathbf{x}|\mathbf{e}) \) is the stationary distribution
- States are instantiations \( \mathbf{x} \) to \( \mathbf{X}-\mathbf{E} \)
- Need transition function that converges to stationary distribution \( P(\mathbf{X}|\mathbf{e}) \)
- For convenience: define \( P_\Phi = P(\mathbf{X}|\mathbf{e}) \) where the factors in \( \Phi \) are reduced by the evidence \( \mathbf{e} \)

Gibbs Sampling Revisited

Using the MCMC framework, the transition model for Gibbs Sampling is:

\[
\mathcal{T}_i((x_{-i}, x_i) \rightarrow (x_{-i}, x'_i)) = P(x'_i | x_{-i})
\]

And the posterior distribution \( P_{\Phi}(\mathbf{X}) = P(\mathbf{X}|\mathbf{e}) \)

is a stationary distribution of this process
Gibbs Sampling Revisited

Gibbs sampling on a Bayesian network is efficient
Note: \( Pa(x_i) = \text{Parents of } x_i \), \( Ch(x_i) = \text{Children of } x_i \)

\[
P(X_i | x_1, ..., x_{i-1}, x_{i+1}, x_n) \frac{P(x_1, ..., x_{i-1}, x_i, x_{i+1}, x_n)}{\sum_{x_i} P(x_1, ..., x_{i-1}, x_i, x_{i+1}, x_n)} = \prod_{j=1}^n P(x_j \mid Pa(x_j))
\]

Depends only on the CPDs of \( X_i \) and its children
Gibbs Sampling Revisited

Block Gibbs Sampling

- Can sample more than a single variable $X_i$ at a time
- Partition $X$ into disjoint blocks of variables $X_1, \ldots, X_k$
- Then sample $P(\Phi(X_i \mid X_1=x_1, \ldots, X_{i-1}=x_{i-1}, X_{i+1}=x_{i+1}, \ldots, X_k=x_k))$
- Takes longer range transitions

Example of Block Gibbs Sampling

Intelligence of 4 students

Difficulty of 2 courses

Grades ($G_{\text{Intelligence, Difficulty}}$)

- Step $t$: Sample all of the $I$ variables as a block, given $D$s and $G$s (since $I$s are conditionally independent from each other given $D$s)
- Step $t+1$: Sample all of the $D$ variables as a block, given $I$s and $G$s (since $D$s are conditionally independent of each other given $I$s)
Gibbs Sampling Revisited

Need to compute $P_{\phi}(X_i \mid X_1 = x_1, \ldots, X_{i-1} = x_{i-1}, X_{i+1} = x_{i+1}, X_k = x_k)$

- Efficient if variables in each block (eg. $I$) are independent given the variables outside the block (eg. $D$)
- In general, full independence is not essential – need some sort of structure to the block-conditional distribution

Gibbs Sampling Revisited

- Gibbs chain not necessarily regular and may not converge to a unique stationary distribution
- Only guaranteed to be regular if $P(X_i \mid X_{i-1})$ is positive for every value of $X_i$
- Theorem 12.4: Let $\mathcal{H}$ be a Markov network such that all of the clique potentials are strictly positive. Then the Gibbs-sampling Markov chain is regular.