

Markov Chain Monte Carlo Methods

Sargur Srihari

srihari@cedar.buffalo.edu

Topics

- Limitations of Likelihood Weighting
- Gibbs Sampling Algorithm
- Markov Chains
- Gibbs Sampling Revisited
- A broader class of Markov chains
- Using a Markov chain
- MCMC in Practice

Limitation of Likelihood weighting

- In likelihood weighting: evidence node affects sampling only for nodes that are descendants
- Effect on non-descendant nodes is accounted for only by the weights
- When the evidence is near the leaf nodes we are essentially sampling from the prior, which is often very far from the desired posterior
- We present an alternative sampling approach that generates a sequence of samples

Sequential sampling

- Sequence is constructed so that
 - although first sample is generated from the prior,
 - successive samples are generated from distributions that get closer to the desired posterior
- Applies equally well to directed and undirected models
- Algorithm is easier to present in terms of factors Φ

Gibbs Sampling Algorithm

- “Fix” the sample by resampling some of the variables we generated early in the process
- Simplest method for doing this is Gibbs sampling presented next
- Start by generating a sample of unobserved variables using some initial distribution
 - use mutilated network and forward sampling
- Iterate over each unobserved variable, sampling a new value for each variable given our current sample for other variables
 - Allows information to flow over the network

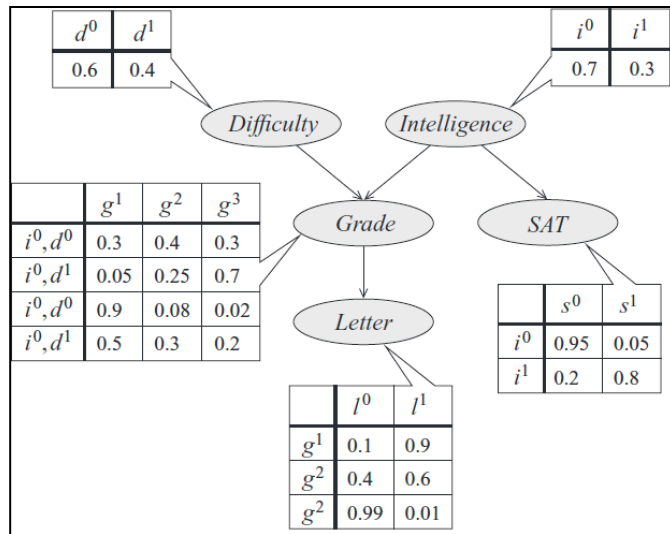
Gibbs Sampling Algorithm

- **Procedure** Gibbs-Sample (
 - \mathbf{X} // Set of variables to be sampled
 - Φ // Set of factors defining P_{Φ}
 - $P^{(0)}(\mathbf{X})$, //Initial state distribution
 - T //Number of time steps)
 - Sample $\mathbf{x}^{(0)}$ from $P^{(0)}(\mathbf{X})$
- **for** $t=1, \dots, T$
 - $\mathbf{x}^{(t)} \leftarrow \mathbf{x}^{(t-1)}$
 - for each** $X_i \in \mathbf{X}$
 - Sample $x_i^{(t)}$ from $P_{\Phi}(X_i | \mathbf{x}_{i-1}^{(t)})$
 - //Change X_i in $\mathbf{x}^{(t)}$
- **return** $\mathbf{x}^{(0)}, \dots, \mathbf{x}^{(T)}$

Gibbs applied to evidence

- To apply to a network with evidence
- We first reduce all of the factors by the observations e so that the distribution P_{Φ} used in the algorithm corresponds to $P(\mathbf{X}|e)$

Ex: Same as one with LW



Evidence: l^0, s^1

1. Algorithm will generate samples over D, I, G
 Set of reduced factors Φ is therefore:
 $P(I), P(D), P(G|I, D), P(s^1|I), P(l^0|G)$

2. Algorithm begins by generating one sample by forward sampling
 Assume this sample is

$$\begin{aligned} d^{(0)} &= d^1 \\ i^{(0)} &= i^0 \\ g^{(0)} &= g^2 \end{aligned}$$

3. We sample unobserved variables D, I, G
 We sample $g^{(1)}$ from $P_\phi(G|d^1, i^0)$
 This computation is efficient (since we are computing the distribution over a single variable given the others)
 Having sampled $g^{(1)} = g^3$ we now continue
 Resampling $i^{(1)}$ from $P_\phi(I|d^1, g^3)$ to get i^1
 Result of first iteration of sampling is:

$$\begin{aligned} d^{(0)} &= d^1 \\ i^{(0)} &= i^1 \\ g^{(0)} &= g^3 \end{aligned}$$

MCMC

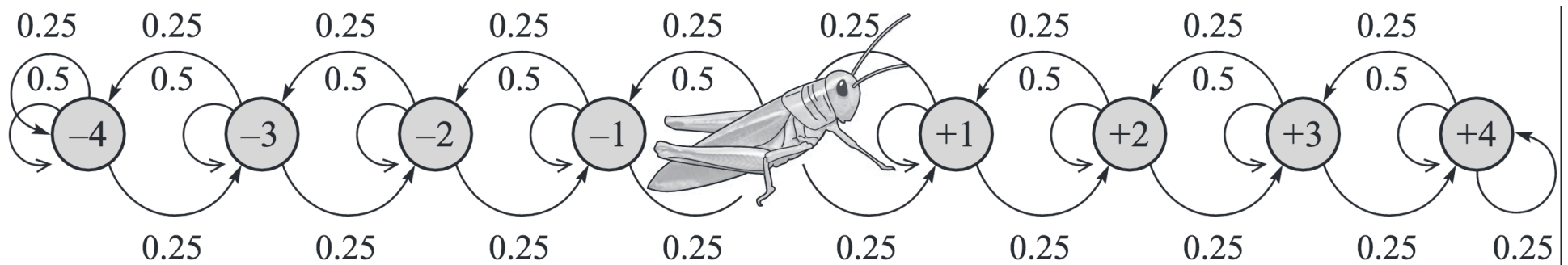
- A general method for generating samples from the posterior distribution

Markov Chain

- Graph is different from PGM
 - It is a graph whose nodes are possible assignments to our variables \mathbf{X}
- A Markov chain is defined via a state-space $Val(\mathbf{X})$ and a model that defines for every state $x \in Val(\mathbf{X})$ a next state distribution over $Val(\mathbf{X})$.
- Transition model \mathcal{T} defines for each pair of states x, x' the probability $\mathcal{T}(x \rightarrow x')$

Ex: Grasshopper Markov Chain

- State consists of nine integers $-4, \dots, +4$ arranged as points on a line. State changes states with probabilities shown



Random Sampling Process

- Defines random state sequence $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots$
 - State of the process at time t is a r.v. $\mathbf{X}^{(t)}$
 - We assume that the initial state $\mathbf{X}^{(0)}$ is distributed according to some initial state distribution $P^{(0)}(\mathbf{X}^{(0)})$
 - We can define distributions over subsequent states $P^{(1)}(\mathbf{X}^{(1)}), P^{(2)}(\mathbf{X}^{(2)}), \dots$ using

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

- Probability of being in state \mathbf{x}' at time $t+1$ is the sum over all possible states \mathbf{x} that the chain could have been at time t of the probability being in state \mathbf{x} times the probability of transition from \mathbf{x} to \mathbf{x}'