COS513 Scribe Notes: Monte Carlo Markov Chain Sampling (II)

Kuk J. Jang

December 1, 2010

1 Overview

- Goal: To build a Markov chain whose stationary distribution is $p(x)$, where $p(x)$ is the target distribution. This can be thought of as approximating the target distribution with an empirical distribution.
- The goal can be thought of as obtaining independent samples from $p(x)$ by adjusting the empirical distribution to match the target distribution. To achieve this we will generate sample from a Markov chain for a "long time" (in relation to the "burn-in") and collect samples at some lag. Intuitively, the lag depends on the "time/number of iterations" before the empirical distribution reaches the stationary distribution, and thus will vary depending on the initial position.

The question that can be asked is when does the Markov chain begin to obtain samples from the stationary distribution? First, a short digression.

2 Definitions and Theorems

First-Order Markov Chains A *First-Order Markov chain* is a series of random variables, $X_1, \ldots, X_m, X_{m+1}$, for which the influence of the values of $X_1, \ldots, X_m$ on the distribution of X_{m+1} depends entirely on X_m (i.e. first-order Markov assumption). Formally,

$$p(X_{m+1}|X_1, X_2, \ldots, X_m) = p(X_{m+1}|X_m) \tag{1}$$

A Markov chain is specified by:

- $p_o(X)$ — the *initial probabilities* of the various states.
- The *transition probabilities* $T_m(x_m, x_{m+1})$ where,

$$T_m(x_m, x_{m+1}) \triangleq p(X_{m+1}|X_m) \tag{2}$$

  where T_m is the probability for state x_{m+1} at time $m+1$ to follow state x_m at time m. *Note: The latent variables of Hidden Markov Models seen in earlier lectures is a Markov chain.*

A Markov chain is ***homogenous*** if the transition probabilities are not a function of time —

$$T_m = T \tag{3}$$

For simplicity, we assume that X is discrete. Using the transition probabilities, one can find the probability of state x occurring at time $m+1$, denoted $p_{m+1}(x)$, from the corresponding probabilities at time m, as follows:

$$p_{m+1}(x_{m+1}) = \sum_{x_m} p_m(x_m) p_(x_{m+1}|x_m) \tag{4}$$

Recall from HMMs that given the initial probabilities, (4) defines a recursion from which we can calculate all the marginal probabilities throughout the Markov chain.

Stationary Distribution A distribution is *stationary* or (*invariant*) with respect to a Markov chain if each step in the chain leaves the distribution given by the probabilities, $p^*(x)$, unchanged for all time, as follows:

$$p^*(x) = \sum_{x'} T(x', x)p^*(x') \tag{5}$$

the marginal over all possible transitions from state x' to state x.

In general, a Markov chain can have 0 or more stationary distributions. A finite Markov chain always has at least one invariant distribution[1].

A *sufficient* condition for stationarity (i.e. implies $p^*(x)$ is a stationary distribution), using notation from before, is as follows:

$$p^*(x)T(x, x') = p^*(x)T(x', x) \tag{6}$$

This is also called the condition of *reversibility* or *detailed balance*.

Proof. Let us check stationarity over detailed balance.

$$\begin{aligned}\sum_{x'} T(x', x)p^*(x') &= \sum_{x'} T(x, x')p^*(x) \ (\textit{From } (6)) \\ &= p^*(x)\sum_{x'} T(x', x) \ (\sum_{x'} T(x', x) = 1) \\ &= p^*(x)\end{aligned}$$

□

It seems logical that in order to achieve our goal of sampling from the target distribution, $p(x)$, we simply have to arrange a Markov chain such that the stationary distribution will be $p(x)$. However, we need an additional condition that as $m \to \infty$, the probabilities at time m, $p_m(x) \to p^*(x)$, regardless of the choice of initial probabilities $p_o(x)$. This property relates to the uniqueness of the invariant distribution of the Markov chain. Moreover, some Markov chains are *periodic*,meaning they "converge" to a cycle of distributions. In order to satisfy this condition, called *ergodicity*, and also determine a bound on the rate of convergence, we turn to the following theorem:

Fundamental Theorem (From Neal, 1993) If a homogeneous Markov chain on a finite space with transition probabilities $T(x, x')$ has $p^*(x)$ as an invariant distribution and

$$\nu = \min_x \min_{x':p^*(x')>0} \frac{T(x, x')}{p^*(x')} > 0 \tag{7}$$

then the Markov chain is *ergodic*, i.e. regardless of the initial probabilities, $p_o(x)$.

$$\lim_{n\to\infty} p_n(x) = p^*(x) \tag{8}$$

A bound on the rate of convergence is given by

$$|p^*(x) - p_n(x)| \leq (1 - \nu)^n \tag{9}$$

This theorem can be interpreted in terms of the state space defined by the model. If transition probabilities allow "escaping" from one state to any other state in the state space of the model, then this theorem guarantees that a unique stationary distribution will exist. Note, however, that this condition is not necessary for ergodicity. The proof of the theorem was not covered during lecture and can be found in the reading material for today's lecture.

[1]From Neal(1993), "Probabilistic Inference Using Markov Chain Monte Carlo Methods"

Computational Effort for Monte Carlo Estimation Recalling that our main goal was to construct a Markov chain with stationary distribution is $p^*(x)$, our desired target distribution, there are several factors that need to be considered in terms of the computational needed:

- The amount of computation needed to simulate a transition from one state to another
- The amount of "time" needed for the chain to converge (i.e. the "burn-in")
- The number of draws need to move from one sample(state) from $p^*(x)$ to another independent sample(state) from $p^*(x)$

For all of these factors there is no dominant theoretical solution. In other words, the theory is still "young." Discussion of these issues can be found in the reading. For now, let's put these issues aside and return to our main goal.

3 The Metropolis-Hastings Algorithm

The idea of the Metropolis algorithm was to draw a sample from some *proposal distribution* and accept the sample according to a criterion, which can be random. Hastings (Hastings, 1970) extended the original Metropolis Algorithm to a more general case.

First, some notation:

- Let the state space of X have K components, $X = X_1, X_2, \ldots, X_K$
  - *e.g. there are K nodes in a graphical model*
- Consider K transition matrices B_k, where each one holds x_j fixed for $j \neq k$.
  - Only the k'th component is changed through a transition from the current state to the next state by B_k
- To move from the current state at time t, $X^{(t)}$, to the next state at time $t+1$, $X^{(t+1)}$, iteratively apply each B_k
  - If detailed balance holds for each B_k, then it will also hold for the product of all the B_k.

Algorithm Suppose x is the current state.

1. Select a candidate state, x' picked at random from the proposal distribution, which may depend on x, given by the probability $B_k(x, x')$.
2. Accept the candidate state as the next state x', with probability $A(x, x')$, the acceptance function. Otherwise, reject it and retain the current state as the next state.

This acceptance function is defined as follows:

$$A_k(x, x') = \min\left(1, \frac{P(x')B_k(x', x)}{P(x)B_k(x, x')}\right) \tag{10}$$

where $P(X)$ is our original target distribution. Note that when the B_k satisfy the symmetry condition $B_k(x', x) = B_k(x, x')$, the algorithm becomes the original Metropolis algorithm.

It remains to be verified that detailed balance is satisfied by this acceptance function.

Proof. For the simple case of one B_k. From (10):

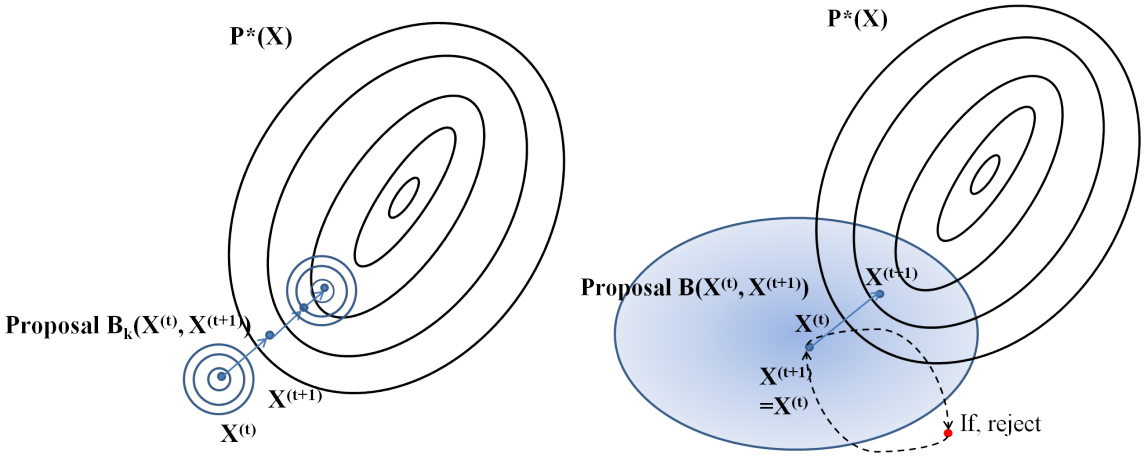
$$\begin{aligned} P(x)B_k(x, x')A_k(x, x') &= \min\left(P(x)B_k(x, x'), P(x')B_k(x', x)\right) \\ &= \min\left(P(x')B_k(x', x), P(x)B_k(x, x')\right) \\ &= P(x)B_k(x', x)A_k(x', x) \end{aligned}$$

Detailed balance is satisfied. If B_k satisfies detailed balance, then the product of all the B_k's will also satisfy detailed balance. Therefore, $P(x)$ is the invariant distribution of the Metropolis-Hastings algorithm. □

4 Additional Remarks

The Metropolis-Hastings algorithm can be applied in many different forms to a sample a target distribution. One way is treating each of the coordinates independently in a "local" algorithm and iteratively apply each of the B_k or choose a random B_k at every iteration. The local algorithm can be illustrated as shown in Figure 1a when applied to a target distribution of a multivariate Gaussian, $P^*(X)$. The proposal distribution $B_k(x, x')$ is limited within the state space. The path of samples follow random walk in the state space, with a tendency to move towards a region with higher probability in the target distribution. Though we are obtaining individual samples, these samples are very correlated and therefore, depending on the condition, may converge very slowly towards the target distribution.

An alternative "global" form is to apply the changes to coordinates at the same time according to a pre-defined proposal distribution, B. This is shown in Figure 1b. In this case, calculating $B(X^t, X^{t+1})$ can be considered applying all B_k. There is a trade-off in this approach as the samples obtained are less correlated and the exploration of the entire state space may be faster, but there is a possibility that more samples will be rejected. To retain reasonable rejection rates, smaller steps may need to be taken in the exploration of the state space compared to the local algorithm. Additional discussion about this issue can be found in the reading in Section 4.4. Figure 1b also illustrates that if a candidate state is rejected, the current state becomes the next state.



(a) Local Algorithm (b) Global Algorithm

Figure 1: Operation of local and global algorithm for a multivariate Gaussian

5 Next Time:

Next time we will be covering the Gibbs sampling algorithm!