Metropolis-Hastings and Gibbs Sampling

Guy Lebanon

November 30, 2006

Markov Chain Monte Carlo Basics

In an earlier note, we saw how samples can be used to approximate expectations

$$\frac{1}{m}\sum_{i=1}^{m} g(x_i) \approx \mathsf{E}_p(g(X)) \quad \text{where} \quad x_1,\ldots,x_m \sim p.$$

We also saw a number of techniques for producing samples from a distribution p such as the histogram and transformation methods and rejection and importance sampling. Markov chain Monte Carlo (MCMC) is a collection of sampling methods that are based on following random walks on Markov chains.

Homogenous Markov chains $X_0, X_1, X_2, \ldots$ are random processes that are completely characterized by the transition probabilities $P(X_n = y|X_{n-1} = z) = T(z, y)$ and initial probabilities $\pi_0(z) = P(X_0 = z)$. To simplify the notation we will assume that X_i are discrete and finite $X_i \in \{1,\ldots,k\}$ and we will consider π and T as a (row) vector and matrix of probabilities. For homogenous Markov processes conditional distribution of X_n given X_{n-1} is independent of $X_1,\ldots,X_{n-2}$. As a result, we have $P(X_1) = \pi_1 = \pi_0 T$. Similarly, $\pi_k = \pi_0 T^k$ and for large k, π_k tends to a unique stationary distribution π satisfying $\pi T = \pi$ (regardless of π_0) if the Markov chain characterized by T is ergodic. In other words, no matter what is the initial distribution π_0 (or where we start from) the resulting position distribution after k steps tends to the stationary distribution π for large k. The idea of MCMC is to generate a random sample from p by following a random walk of k steps on a Markov chain T, for which p equals its stationary distribution π.

Thus, no matter where we start, if we follow a random walk for a long period (called burn-in time) we will end up with a sample from its stationary distribution. If we want several samples, we can either (i) repeat the process several times (ii) take consecutive samples after the burn in time or (iii) follow a random walk and record every l-step as a sample. Approach (ii) will not produce independent samples and approach (iii) will result in approximately independent samples from π if l is sufficiently large.

To sample from p using MCMC, we need to design a Markov chain T whose stationary distribution π is p. To ensure that, it suffices to show that T satisfies the detailed balance property with respect to p

$$p_i T_{ij} = p_j T_{ji} \quad \forall\, i, j$$

since then

$$[pT]_i = \sum_j p_j T_{ji} = \sum_j p_i T_{ij} = p_i \sum_j T_{ij} = p_i \quad \Rightarrow \quad pT = p.$$

We also need to ensure that the Markov chain described by T is ergodic so there will be a unique stationary distribution. One simple way to ensure ergodicity of T is to have $T_{ij} > 0$ for all i, j. It is useful to know (and easy to verify) that if we have several Markov chains $T_1,\ldots,T_l$ that satisfy the detailed balance property then a linear combination of them $\sum_i \alpha_i T_i$ would also satisfy it.

The Metropolis-Hastings Algorithm

The Metropolis-Hastings sampling model constructs an ergodic Markov chain that satisfies the detailed balance property with respect to p and therefore produce the appropriate samples. The transition T is

based on sampling from a proposal conditional distribution $q(z|z^{(t)})$ (which we assume may be easily done). Specifically, given the t-step in the random walk $z^{(t)}$ we generate the next step $z^{(t+1)}$ as follows:

$$z^{(t+1)} = \begin{cases} z' & \text{with probability } r(z^{(t)}, z') = \min\left(1, \frac{p(z')}{p(z^{(t)})} \frac{q(z^{(t)}|z')}{q(z'|z^{(t)})}\right) \\ z^{(t)} & \text{with probability } 1 - r(z^{(t)}, z') \end{cases}$$

where $z' \sim q(z|z^{(t)})$. The two stage process results in the following Markov transition

$$T(z^{(t)}, z^{(t+1)}) = r(z^{(t)}, z^{(t+1)}) q(z^{(t+1)}|z^{(t)}) + \left(1 - \sum_{z'} r(z^{(t)}, z') q(z'|z^{(t)})\right) \delta_{z^{(t)}, z^{(t+1)}}.$$

T is ergodic if $q(z|z^{(t)}) > 0$ and detailed balance w.r.t p holds since T above is written as a sum of two matrices that satisfy the detailed balance property w.r.t p

$$p(z) r(z, z') q(z'|z) = \min\left(p(z) q(z'|z), p(z) q(z'|z) \frac{p(z')}{p(z)} \frac{q(z|z')}{q(z'|z)}\right) = \min(p(z) q(z'|z), p(z') q(z|z')) = p(z') r(z', z) q(z|z')$$

$$p(z) \left(1 - \sum_{z'} r(z, z') q(z'|z)\right) \delta_{z,z'} = p(z') \left(1 - \sum_{z} r(z', z) q(z|z')\right) \delta_{z',z}$$

In practice, the proposal is often taken to be a Gaussian $q(z'|z) = N(z'; z, \sigma^2 I)$ which is an easy distribution to sample from (for example using the Box-Müller method[1]). In this and related other case $q(z'|z) = q(z'|z)$ and the acceptance probability simplifies to $\min(1, \frac{p(z')}{p(z^{(t)})})$ which demonstrates that if the proposed state is more likely than the old one, it is accepted with probability 1. If the proposed state z' is less likely than the current one $z^{(t)}$, the probability of accepting depends on the likelihood ratio $p(z')/p(z^{(t)})$. Choosing a proposal with small variance (for example $\sigma^2 \to 0$ for the above Gaussian proposal) would result in relatively high acceptance rates but with strongly correlated consecutive samples. Increasing the variance would de-correlate consecutive accepted samples to some extent, but it is also likely to reduce the acceptance rate.

Gibbs Sampling

Gibbs sampling is a special case of Metropolis-Hastings where the proposal q is based on the following two stage procedure. First, a single dimension i of z is chosen randomly (say uniformly). The proposed value z' is identical to z except for its value along the i-dimension z_i is sampled from the conditional $p(z_i|z_{-i}^{(t)})$ where $z_{-i}^{(t)} = \{z_1^{(t)}, \ldots, z_{i-1}^{(t)}, z_{i+1}^{(t)}, \ldots, z_m^{(t)}\}$. Since

$$\frac{p(z')}{p(z^{(t)})} \frac{q(z^{(t)}|z')}{q(z'|z^{(t)})} = \frac{p(z_i'|z_{-i}') p(z_{-i}')}{p(z_i^{(t)}|z_{-i}^{(t)}) p(z_{-i}^{(t)})} \frac{p(z_i^{(t)}|z_{-i}')}{p(z_i|z_{-i}^{(t)})} = \frac{p(z_i'|z_{-i}') p(z_{-i}')}{p(z_i^{(t)}|z_{-i}') p(z_{-i}')} \frac{p(z_i^{(t)}|z_{-i}')}{p(z_i|z_{-i}')} = 1$$

the acceptance rate is always 1 and Gibbs sampling performs a random walk where at each iteration the value along a randomly selected dimension is updated according to the conditional distribution (geometrically, this constitutes axis aligned transitions). The detailed balance property holds since Gibbs is a special case of Metropolis and T is ergodic if all dimensions are updated with positive probability.

Gibbs sampling is useful when sampling from $p(z_i|z_{-i}^{(t)})$ is easy and quick. In these cases, each random walk iteration is quick and all proposed values are accepted. Examples for such models are Bayesian networks or other models that are specified as a product of conditional distributions.

[1] See for example Numerical Recipes in C, `http://www.nrbook.com/a/bookcpdf/c7-2.pdf`