Notes on Metropolis Hastings

Matthias Seeger
Department of EECS
University of California at Berkeley
485 Soda Hall, Berkeley CA
mseeger@cs.berkeley.edu

August 5, 2005

1 The Metropolis Hastings Algorithm

Let $\pi(\boldsymbol{x})$ be the density of a distribution we would like to draw samples from. A *Markov Chain Monte Carlo (MCMC)* method does this by running a Markov chain with a transition kernel $T(\boldsymbol{x}^*|\boldsymbol{x})$ (being a conditional probability density) which leaves π invariant, in the sense that if $\boldsymbol{x} \sim \pi$, $\boldsymbol{x}^* \sim T(\cdot|\boldsymbol{x})$, then $\boldsymbol{x}^* \sim \pi$. If the chain is also ergodic and nonperiodic (see relevant literature), the marginal distributions of variables down the chain will converge to π.

A simple sufficient condition for invariance is *detailed balance*:

$$T(\boldsymbol{x}^*|\boldsymbol{x})\pi(\boldsymbol{x}) = T(\boldsymbol{x}|\boldsymbol{x}^*)\pi(\boldsymbol{x}^*) \quad \text{for all } \boldsymbol{x},\ \boldsymbol{x}^*.$$

The Metropolis Hastings procedure is a general way of constructing kernels which fulfil the detailed balance condition. Let $q(\boldsymbol{x}^*|\boldsymbol{x})$ be an arbitrary proposal distribution and define

$$\alpha(\boldsymbol{x}^*, \boldsymbol{x}) = \min\left\{1, \frac{q(\boldsymbol{x}|\boldsymbol{x}^*)\pi(\boldsymbol{x}^*)}{q(\boldsymbol{x}^*|\boldsymbol{x})\pi(\boldsymbol{x})}\right\}.$$

Then the MH kernel $T_q(\cdot|\boldsymbol{x})$ samples $\boldsymbol{x}' \sim q(\cdot|\boldsymbol{x})$, evaluates $\alpha = \alpha(\boldsymbol{x}', \boldsymbol{x})$ and sets $\boldsymbol{x}^* = \boldsymbol{x}'$ with probability α (acceptance of the proposal $\boldsymbol{x}'$), $\boldsymbol{x}^* = \boldsymbol{x}$ otherwise (rejectance).

In order to prove detailed balance, let

$$A(\boldsymbol{x}', \boldsymbol{x}) = \alpha(\boldsymbol{x}', \boldsymbol{x})q(\boldsymbol{x}'|\boldsymbol{x})\pi(\boldsymbol{x}) = \min\left\{q(\boldsymbol{x}'|\boldsymbol{x})\pi(\boldsymbol{x}), q(\boldsymbol{x}|\boldsymbol{x}')\pi(\boldsymbol{x}')\right\}.$$

Importantly, A is symmetric: $A(\boldsymbol{x}', \boldsymbol{x}) = A(\boldsymbol{x}, \boldsymbol{x}')$. Now,

$$T_q(\boldsymbol{x}^*|\boldsymbol{x})\pi(\boldsymbol{x}) = \int A(\boldsymbol{x}', \boldsymbol{x})\delta_{\boldsymbol{x}^*,\boldsymbol{x}'} + B(\boldsymbol{x}', \boldsymbol{x})\delta_{\boldsymbol{x}^*,\boldsymbol{x}}\, d\boldsymbol{x}' = A(\boldsymbol{x}^*, \boldsymbol{x}) + \kappa(\boldsymbol{x})\delta_{\boldsymbol{x}^*,\boldsymbol{x}}$$

with $B = q\pi - A$. The r.h.s. is symmetric, which proves detailed balance.

The procedure originally suggested by Metropolis *et.al.* was restricted to symmetric proposal distributions in which case the acceptance probability becomes $\alpha = \min\{1, \pi(\boldsymbol{x}^*)/\pi(\boldsymbol{x})\}$. The clear advantage of using a symmetric proposal distribution is that $q(\boldsymbol{x}^*|\boldsymbol{x})$ and $q(\boldsymbol{x}|\boldsymbol{x}^*)$ do not have to be computed explicitly. If the proposal is not symmetric, but is still hard to compute, a random MH scheme might work. This is the topic of the next Section.

1.1 Random Choice of Proposal Distribution

Suppose there is a family $q(\cdot|\boldsymbol{x},\beta)$ of proposal distributions (the family is indexed by β which is a parameter independent of $\boldsymbol{x}$) and we would like to choose one of them at random according to $P(\beta|\boldsymbol{x})$. In general we can use the MH kernel with the marginal proposal distribution $\int q(\cdot|\boldsymbol{x},\beta)P(\beta|\boldsymbol{x})\,d\beta$, but this marginal may be hard to compute. If it is true that

$$q(\boldsymbol{x}'|\boldsymbol{x},\beta)P(\beta|\boldsymbol{x})\pi(\boldsymbol{x}) = q(\boldsymbol{x}'|\boldsymbol{x},\beta)P(\beta|\boldsymbol{x}')\pi(\boldsymbol{x})$$

for all $\boldsymbol{x}$, $\boldsymbol{x}'$, β, we can also simply sample β along with $\boldsymbol{x}'$ and use the conditional acceptance probability $\alpha(\boldsymbol{x}',\boldsymbol{x},\beta)$ where β is simply plugged in. To prove detailed balance, define $A(\boldsymbol{x}',\boldsymbol{x},\beta) = \alpha(\boldsymbol{x}',\boldsymbol{x},\beta)q(\boldsymbol{x}'|\boldsymbol{x},\beta)\pi(\boldsymbol{x})$, and note that $A(\boldsymbol{x}',\boldsymbol{x},\beta)P(\beta|\boldsymbol{x})$ is symmetric for every β by the assumption on $P(\beta|\boldsymbol{x})$. Therefore, the marginal $\tilde{A}(\boldsymbol{x}',\boldsymbol{x}) = \int A(\boldsymbol{x}',\boldsymbol{x},\beta)P(\beta|\boldsymbol{x})\,d\beta$ is symmetric, which implies detailed balance. The assumption on $P(\beta|\boldsymbol{x})$ is of course satisfied if $P(\beta|\boldsymbol{x})$ does not depend on $\boldsymbol{x}$. A more interesting and practically useful case is that $P(\beta|\boldsymbol{x}) = P(\beta_2|f(\boldsymbol{x},\beta_1))P(\beta_1)$ for $\beta = (\beta_1,\beta_2)$ and some mapping f such that whenever a transition from $\boldsymbol{x}$ to $\boldsymbol{x}'$ is possible under $q(\cdot|\boldsymbol{x},\beta)$, $\beta \sim P(\cdot|\boldsymbol{x})$, then $f(\boldsymbol{x},\beta_1) = f(\boldsymbol{x}',\beta_1)$.

An example of this setup can be found in [1]. There, $\boldsymbol{x}$ consists of a partition of n observations into groups. $\boldsymbol{x}$ is to be updated using split and merge steps. For a group picked as a split candidate, the authors suggest to run Gibbs sampling on the group indicators, based on a "posterior" which is restricted to the data within the group. After a while, the new state is taken as MH proposal. Clearly this is not a symmetric proposal, and if each indicator is updated more than once, then it is hard in general to compute the proposal probabilities. To overcome this problem, the authors use random MH, where all Gibbs updates except for a single final run over the group indicators are lumped into a random proposal. In other words, the family of proposals is indexed by states $\hat{\boldsymbol{x}}$ which agree with $\boldsymbol{x}$ outside the group, and $P(\beta|\boldsymbol{x})$ is implemented using restricted Gibbs sampling. States picked in this way are referred to as *launch states*. Using our notation, β_1 is a pair k, l of observations chosen at random, and a split is proposed if these observations belong to the same group (a merge is proposed otherwise). $f(\boldsymbol{x},\beta_1)$ is the set of all observations assigned to the same group(s) than k, l, this set is called S in [1]. If k, l belong to the same group, the proposal is to split the corresponding group. In this case, β_2 is the launch state obtained using restricted Gibbs sampling on the points in $S \setminus \{k,l\}$. The final proposal starts from β_2 for a single restricted Gibbs run over these points, for which the proposal probabilities can be computed easily.[1] If k, l belong to different groups, the proposal is to merge them into a single one. Note that even in this case we have to actually sample β_2 in order to compute the MH acceptance probability, namely because $q(\boldsymbol{x}|\boldsymbol{x}^*,\beta)$ appears in the MH ratio.

Since the launch state β_2 depends on $\boldsymbol{x}$ only through $S = f(\boldsymbol{x},\beta_1)$ which is clearly the same as $S' = f(\boldsymbol{x}',\beta_1)$ (*i.e.* a split or merge for a given β_1 leaves the set S invariant, as long as the same k, l are picked), we can use MH without having to compute the marginal proposal distribution (which would be intractable in the case of [1])[2].

If $P(\beta|\boldsymbol{x})$ does not fulfil the symmetry property and the marginal proposal distribution cannot be computed easily, we have the option of extending the state space to include β

[1] Since there are no latent variables which have to be summed out.

[2] Unfortunately, Jain and Neal [1] do not mention the argument made here in their paper, and the reference they cite for justifying random MH kernels only deals with the trivial case $P(\beta|\boldsymbol{x}) = P(\beta)$.

along with $\boldsymbol{x}$, with an obvious modification of the MH procedure. In this case, β is an *auxiliary variable* which is dragged along for computational reasons (faster mixing, *etc.*) only.

1.2 Gibbs Sampling

Suppose $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2)$, and that we can sample from $\pi(\boldsymbol{x}_j|\boldsymbol{x}_{3-j})$ easily (the latter is called *full conditional distribution*). Then, the *Gibbs kernels* are MH kernels:

$$q^{(j)}(\boldsymbol{x}^*|\boldsymbol{x}) = \pi(\boldsymbol{x}_j^*|\boldsymbol{x}_{3-j})\delta_{\boldsymbol{x}_{3-j}^*,\boldsymbol{x}_{3-j}},\ j = 1, 2.$$

Note that if $q^{(j)}$ is used, then $\alpha(\boldsymbol{x}^*, \boldsymbol{x}) = 1$ if $\boldsymbol{x}_{3-j}^* = \boldsymbol{x}_{3-j}$, so all Gibbs proposals are accepted. The generalization to more than two blocks of $\boldsymbol{x}$ is straightforward. In order to ensure egodicity, one has to cycle through all different j using some scheme which visits each block infinitely often (this is necessary, but not sufficient).

References

[1] S. Jain and M. Neal, R. A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. Technical Report 2003, Department of Statistics, University of Toronto, July 2000.