Explanatory Notes 4 for Bayesian Inference

The Metropolis-Hastings algorithm

Takis Konstantopoulos, Spring 2009

What, perhaps, is not clear to you in the notes (e.g. SZ pp.37-onwards) is the rationale for the algorithm and that its most essential element is often the complexity of the set S rather than the probability π on S. The theory is trivial. It is its applications in specific cases that generates a lot of work (and interest).

PROBLEM: We are given a probability π on some "complicated" finite set S.

Here is an example of a complicated set. Consider a sudoku puzzle. Usually, it has a unique solution. Imagine that I erase half the numbers that are already there. Then there are many solutions. Let S be the set of solutions. We would like, for instance, to produce one solution at random, meaning to be chosen from a uniform probability distribution on S. Clearly, enumerating S is a complex problem. Can we simulate a uniform distribution on S without knowing the size of S (even if we suspect that this size is unimaginably huge)?

ANSWER: Yes we can, by creating a Markov chain whose stationary (and limiting) distribution is π.

Here is a method.

First define a neighbourhood structure on S. This means: for each $x \in S$ decide which elements of S are neighbours of x. We want this set $N(x)$ to be relatively small. It is chosen in a way that is dictated by the definition of S.

Example: Let S be the set of permutations of n objects. Pick a permutation $x = (x_1, \ldots, x_n)$. (Hence the $x_1, \ldots, x_n$ are distinct numbers.) Define another permutation y of x to be a neighbour of it if $y_i = x_i$ for all i except two values. For instance, with $n = 6$, the permutation $x = (2, 4, 1, 6, 5)$ has $(4, 2, 1, 6, 5)$, $(1, 4, 2, 6, 5), \ldots, (2, 4, 1, 5, 6)$ as neighbours.

We require the neighbourhood structure to be symmetric: if y is a neighbour of x then x is a neighbour of y. In other words, $y \in N(x) \iff x \in N(y)$.

Once we have a neighbourhood structure, we can define a way to move on the set S, respecting this structure. (Incidentally, a neighbourhood structure is equivalent to making S a graph, i.e. to declare that certain pairs of states $\{x, y\}$ as neighbours by joining them by an edge. So the set of edges E is the set of $\{x, y\}$ such that $y \in N(x)$.) The way we propose a move is, *for*

example, by picking one of the neighbouring states at random. So, when we are at x, we decide to pick as next state y with probability

$$q(x,y) = \frac{1}{|N(x)|}, \quad y \in N(x).$$

It is NOT necessary to pick one of the neighbours at random. In fact, all we require for the method to work in theory is a set of transition probabilities $q(x,y)$, i.e.

$$q(x,y) \geq 0, \quad \sum_y q(x,y) = 1,$$

such that

$$q(x,y) > 0 \iff q(y,x) > 0.$$

Clearly, we can define a Markov chain with transition probabilities $q(x,y)$. But this is NOT a Markov chain that has, in general, stationary distribution π. (Why should it? The $q(x,y)$ are chosen rather arbitrarily, simply respecting the logical structure of S.) Instead, we modify the $q(x,y)$ to obtain a chain with the required property that its stationary distribution is π.

Pick a function $h : (0,\infty) \to (0,1]$ with the property

$$h(u) = uh(1/u), \quad u > 0.$$

Examples of such a function are:

$$h(u) = \min(1,u),$$

or

$$h(u) = \frac{u}{u+1}.$$

Next let

$$\alpha(x,y) := \begin{cases} h\left(\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right), & \text{if } q(x,y) > 0, \\ 0 & \text{if not.} \end{cases}$$

Let

$$p(x,y) := q(x,y)\alpha(x,y), \quad x \neq y.$$

(And let $p(x,x) = 1 - \sum_{y \neq x} p(x,y)$.)

Theorem 1. *The Markov chain on S with transition probabilities $p(x,y)$ is a reversible Markov chain with stationary distribution π.*

Proof. First observe that h has the property

$$vh(\frac{u}{v}) = uh(\frac{v}{u}).$$

We claim that the detailed balance equations

$$\pi(x)p(x,y) = \pi(y)p(y,x)$$

are satisfied for any pair of states x, y such that $q(x,y) > 0$. Indeed, the display above is written as

$$\pi(x)q(x,y)h\left(\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right) = \pi(y)q(y,x)h\left(\frac{\pi(x)q(x,y)}{\pi(y)q(y,x)}\right),$$

which is an immediate consequence of the aforementioned property of h. □

SIMULATION: In practice, the way we simulate the transition probabilties $p(x,y) = q(x,y)\alpha(x,y)$ is by taking into account that $q(x,y)$ are also transition probabilties and that $0 < \alpha \leq 1$.

If the chain is currently in state x then the next state (call it Y) must be chosen such that $P(Y = y) = p(x,y)$.

Do this in two steps:

First choose a random variable Z with

$$P(Z = y) = q(x,y).$$

Next toss a coin, that is, a random variable $\xi \in \{0,1\}$ such that

$$P(\xi = 1|Z = y) = \alpha(x,y)$$
$$P(\xi = 0|Z = y) = 1 - \alpha(x,y).$$

Finally, let

$$Y := \begin{cases} Z, & \text{if } \xi = 1, \\ x, & \text{otherwise.} \end{cases}$$

It is easy to see that $P(Y = y) = p(x,y)$. Indeed, if $y \neq x$,

$$P(Y = y) = P(Z = y, \xi = 1) = P(Z = y)P(\xi = 1|Z = y) = q(x,y)\alpha(x,y) = p(x,y).$$

This is why we call $q(x,y)$ as PROPOSAL probabilities, while $\alpha(x,y)$ as ACCEPTANCE probabilties.

The logic of writing the code is now clear:

Generate a proposed next state y according to $q(x, y)$.
Toss a coin with probability of heads $\alpha(x, y)$.
If heads come up, then accept the next state and let y be the actual next state of the chain.
Otherwise, if tails come up, do not accept the state y and let the chain remain at state x for another step.
Repeat.

Exercise Consider a $N \times N$ grid G, represented by the pairs of numbers (i, j), $1 \leq i, j \leq N$. Call each such pair a "site". On each site place a ball with one of two colours: colour 1 or colour 0. A configuration of the system is the balls with their colours. However, we do not wish to consider certain configurations. In fact, we do not like configurations in which two neighbouring balls have colour 1. (If a ball is in site (i, j) then its neighbouring balls are the ones in sites $(i \pm 1, j \pm 1)$, i.e. up/down/left/right.) Call such a configuration an acceptable configuration. We let S be the set of acceptable configurations. For example, with $N = 4$,

$$\begin{array}{cccc} 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{array}$$

is not an acceptable configuration, but

$$\begin{array}{cccc} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{array}$$

is. Now let π be the uniform probability on S.

To simulate from π, we first define a neighbourhood structure on S. Thus, given an acceptable configuration, such as in the last display, how would you choose its neighbouring configurations? Can you choose them in a way that not two many balls differ in colour? Can you do it so that the resulting graph is irreducible?

Once you're done with this "logical" step, then you are essentially done and you can write the R code.

Run the Markov chain for $n = 10000$ steps. Then X_n will be approximately uniform on S, i.e.

$$P(X_n = x) \approx 1/|S|,$$

where $|S|$ is the size of $|S$. Now you can use this to estimate the size of S.

THE ALGORITHM ON NON-DISCRETE SETS: Suppose that we want to simulate from a DENSITY $\pi(x)$, and (for simplicity!) let $x \in \mathbb{R}$. So $\pi(x) \geq 0$ and

$$\int_{\mathbb{R}} \pi(x)dx = 1.$$

Consider any other density f you like and can simulate:

For example, say f is uniform on $(-\delta, \delta)$, for some positive number δ.

As another example, take f to be the density of a normal random variable with mean 0 and variance δ.

Define proposal transition densities by

$$q(x, y) := f(y - x).$$

(This means that we, implicitly, consider a Markov chain Z_n with $P(Z_{n+1} \in dy|Z_n = x) = q(x, y)dy$,) Now define acceptance probabilities

$$\alpha(x, y) = h\left(\frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right)$$

(If f is a symmetric density arounf 0 then this formula simplifies: the q-terms cancel.) The actual transition densities are defined to be

$$p(x, y) = q(x, y)\alpha(x, y).$$

Thus, the Markov chain we create (using the coin-tossing method as in the discrete space case) has

$$P(X_{n+1} \in dy|X_n = x) = p(x, y)dy.$$

You can check that

$$\pi(x)p(x, y) = \pi(y)p(y, x),$$

and this will imply that if the Markov chain (X_n) is started with $P(X_0 \in dx) = \pi(x)dx$ then $P(X_n \in dx) = \pi(x)dx$ for all $n \geq 1$. Moreover, if the Markov chain is started from any X_0 (which does not depend on the future), then the density of X_n will converge to $\pi(x)$.