Abstract

Bringing Metropolis to Statistics, Hastings generalization, Component-wise Metropolis, Gibbs Sampler.

Connection with Statistics

To sample from an arbitrary random vector $X \in R^p$ with pdf f just turn f into a canonical distribution by choosing T and $Z(T)$ and defining an energy function by,

$$E(x) = -T \log\ f(x) - T \log\ Z(T)$$

so that,

$$f(x) = \frac{1}{Z(T)} \exp\left\{-\frac{E(x)}{T}\right\}$$

Usually we only know $f(x)$ up to a proportionality constant independent of x, i.e.

$$f(x) = \frac{1}{c} h(x) \text{ with } \ c = \int h(x) dx$$

$h(x)$ is assumed to be known and computable but c may be unknown. In this case we take $T = 1$ and,

$$E(x) = -\log\ h(x)$$

as the energy function. This is ok for sampling from f. To compute the mode of f, we can run the simulated annealing algorithm with the above energy but with a general value for T.

Plain Metropolis with Symmetric Proposal Distribution

The plain vanilla Metropolis algorithm generates a stochastic sequence of states, from an arbitrary starting vector x_0, and the following rule to go from a current state x to a new state y:

1. Propose a candidate state y from a proposal distribution, $q(y|x)$ that may depend on the current x. This proposal distribution is assumed to be symmetric in x and y, i.e., for plain vanilla Metropolis,

   $$q(y|x) = q(x|y)$$

   this can be easily implemented, e.g., take

$$y = x + \epsilon$$

with ϵ a random vector radially symmetric about 0. Typical choices for the distribution of ϵ are multivariate Cauchy, Gaussian, or T distribution with a few degrees of freedom.

2. If the new proposed state y uses less energy than the current state x then go there with probability one. If the new state is more expensive, in terms of energy, than the one we are currently on, then test your luck and go there with a probability exponentially decreasing in the difference of energy. More formally the acceptance probability is,

$$\begin{aligned} \alpha(x, y) &= \exp\left\{-\frac{(E(y) - E(x))_+}{T}\right\} \\ &= \min\{1, \frac{f(y)}{f(x)}\} \end{aligned}$$

The Plain Vanilla Metropolis

```
{
X_0 <- x_0 (arbitrary)
for t = 0,1,2,...

    {
    y <-  sample from q(.|x_t)
    u <- unif(0,1)
    if u <= min{1, f(y)/f(x_t)} then X_{t+1} <- y
    else X_{t+1} <- x_t
    }

  }
```

Example: Javascript implementation of plain vanilla Metropolis for sampling from $N(0, 1)$ with

$$q(y|x) \equiv \text{unif}(x - 1, x + 1)$$

Solution by Haihong Li

(*http://omega.stat.psu.edu:8008/summer99/lecture3/haihong-1.html*)

Metropolis-Hastings

A simple modification to the acceptance probabilities used in the plain vanilla algorithm, allows to use non-symmetrical proposal distributions and still have detailed balance. Change the previous formula for α to,

$$\alpha(x,y) = \min\left\{1, \frac{q(x|y)f(y)}{q(y|x)f(x)}\right\}$$

All we need to do is to check detailed balance, i.e.,

$$p(y|x)f(x) = p(x|y)f(y) \text{ for all } x, y$$

But this is straight forward to check. When $x = y$ it is obviously true, and for $x \neq y$ the only way to arrive to y is by accepting it as a proposed candidate. Thus, the above equation just says that,

$$\alpha(x,y)q(y|x)f(x) = \alpha(y,x)q(x|y)f(y)$$

which is the same as,

$$\min\left\{1, \frac{q(x|y)f(y)}{q(y|x)f(x)}\right\} q(y|x)f(x) = \min\left\{1, \frac{q(y|x)f(x)}{q(x|y)f(y)}\right\} q(x|y)f(y)$$

which is true, since when the min. on one side is 1, the min. on the other side isn't and the denominator cancels with the term outside the parenthesis producing the equality of both sides.

As always, detailed balance is enough to assure that f is the stationary distribution for the chain. But detailed balance is not necessary. It is possible for a chain to have f as its stationary distribution without $p(y|x)f(x) = p(x|y)f(y)$ for all x, y. All that it is needed for stationarity is, (its definition)

$$f(x) = \int p(x|y)f(y)dy$$

i.e., that f be the fix point for the linear operator T, defined by the transition kernel as,

$$T : f(\cdot) \to \int p(\cdot|y)f(y)dy$$

Knowing that f is a stationary distribution for a Markov Chain with transition kernel $p(x|y)$ is not enough to warrant that the chain will eventually sample states according to the density f. It only says that if at some time t the density of X_t is f then it will remain f for all subsequent times since the law of X_{t+1} is $Tf = f$. From the classical fix point theorem of functional analysis we know that if the f's belong to a complete metric space, with distance function d, and the operator T is contractive, i.e.

$$d(Tf, Tg) \leq \lambda d(f, g) \text{ with } \lambda < 1$$

then the iterates $f_0, T(f_0), T(T(f_0)), \ldots$ converge geometrically in the metric of the space to the unique fix point for T, provided that the initial f_0 is not too far from the fix point. Notice that,

$$\text{if } X_0 \text{ comes from } f_0 \text{ then } X_t \text{comes from } T^t f_0$$

Thus, the Markov Chain is expected to have the fix point as its long-run distribution. As we will see later the conditions of the classic fix point theorem are too restrictive. Under mild regularity conditions on the kernel (e.g. when T is only contractive on the average) a large class of Markov Chains will be geometrically ergodic. More on this theme when we look at convergence theorems...

Component-Wise Metropolis

When the dimensionality of the random vector X is large, it is often computationally more efficient to separate the coordinates of X into groups and use proposal distributions that update only one group of coordinates at a time. Towards this end, we introduce the following notation. Let $I = \{1, 2, \ldots, p\}$ be the set of indices for the coordinates of x. We write,

$$x = \{x^1, x^2, \ldots, x^p\} \equiv x^I$$

If $J \subset I$ we write x^{-J} the set of all coordinates except those in the set J, i.e. $x^{-J} = x^{I \setminus J}$. Let $I_1, I_2, \ldots, I_m$ form a partition of I so that x gets separated into m groups,

$$x = \{x^{I_1}, \ldots, x^{I_m}\}$$

Start the Metropolis-Hastings algorithm from somewhere and suppose that at time t we are visiting state x, then at time $t+1$ we either remain at x or we go to y by modifying only one of the m components of x, i.e.,

$$x_{t+1} = y = \{x^{-I_k}, y^{I_k}\} \text{ for some } k$$

The value of k can be chosen at random among $\{1, 2, \ldots, m\}$ or one after another cyclically, in which case to have a homogeneous Markov chain we need to define one iteration only after a full sweep over the set of m components is completed. Choosing the components at random eliminates this problem but may end up, by chance, neglecting some of the components for a long time. An intermediate solution is to randomize the order and then update the components in that order. In all these cases the acceptance probability is the usual Metropolis-Hastings formula but when x differs from y in the k component we have,

$$\min\left\{1, \frac{q(x|y)f(y)}{q(y|x)f(x)}\right\} = \min\left\{1, \frac{q_k(x^{I_k}|y^{I_k}, x^{-I_k})f(y^{I_k}|x^{-I_k})}{q_k(y^{I_k}|x^{I_k}, x^{-I_k})f(x^{I_k}|x^{-I_k})}\right\}$$

where q_k is the proposal function q for updating component k and ratio $f(y)/f(x)$ is written as the ratio of full conditionals, where,

$$f(x^{I_k}|x^{-I_k}) = \frac{f(x^{I_k}, x^{-I_k})}{\int f(y^{I_k}, x^{-I_k})dy^{I_k}}$$

By writing the acceptance probability in this way we can see that the Markov Chain associated to the sequence of updates of component k, with all the other components fix at x^{-I_k}, has the full conditional (defined above) as stationary distribution. Eventually all the components end up sampling from the correct conditionals.

Gibbs Sampler

When the proposal distribution for the k component of x is taken as the full conditional itself we get the Gibbs sampler, i.e. when

$$q_k(y^{I_k}|x^{I_k}, x^{-I_k}) = f(y^{I_k}|x^{-I_k})$$

In this case the acceptance probability becomes 1 and the Gibbs sampler always accepts the proposal candidates. Gibbs sampling is just a special case of Metropolis. This is the justification to the popular statement: *to sample from a joint distribution just sample repeatedly from its one dimensional conditionals given whatever you've seen at the time.*

The Plain Vanilla Gibbs

$\{$
$X_0 \leftarrow x_0$ (arbitrary)
for $t = 0, 1, 2, \ldots$

$\quad\{$
$\quad k \leftarrow$ unif. on $\{1, 2, \ldots, p\}$
$\quad y^k \leftarrow$ sample from $f(x^k|x_t^{-\{k\}})$
$\quad X_{t+1} \leftarrow \{y^k, x_t^{-\{k\}}\}$ $\}$

$\}$

Example: Javascript implementation of Gibbs sampler for generating samples from (X, Y) with one dimensional conditionals given by:

$$\begin{aligned} g_1(x|y) &= (1-y)^{-1} \text{ for } 0 < y < x < 1 \\ g_2(y|x) &= \frac{3y^2}{x^3} \text{ for } 0 < y < x < 1 \end{aligned}$$