Why does the Metropolis-Hastings procedure satisfy the detailed balance criterion?

Kris Hauser

10/21/2013

Given an (unnormalized) target distribution $P(x)$ that is hard to sample from and simple proposal distribution $Q(x'|x)$, the Metropolis-Hastings (MH) algorithm generates a sequence of iterates $x_0, \dots, x_N$ whose distribution, with sufficiently large N, approximate P. However, this fact is rarely proven when the algorithm is presented. This document attempts to shed some light on the subject in a way that is understandable to, say, an upper-level engineering undergraduate or introductory graduate student who has had some basic exposure to the method.

MH is typically stated as follows:

`Metropolis-Hastings`(x_0, P,Q)
For $k = 0,..,N-1$ do:

1. Sample x' from $Q(x'|x_k)$.
2. With *acceptance probability* $\alpha(x', x_k) = \min\left(1, \frac{P(x')Q(x_k|x')}{P(x_k)Q(x'|x_k)}\right)$, set $x_{k+1} \leftarrow x'$.
3. Otherwise, set $x_{k+1} \leftarrow x_k$.

If x_{k+1} is set in Step 2, we call it an *accepted* step, otherwise we call it a *rejected* step.

It is usually stated without proof that as N grows large and with relatively mild restrictions on P and Q, the distribution of $x_0, \dots, x_N$ approaches P. But why? The answer is completely unobvious on a first, second, or even third glance. The acceptance probability is mysterious, with its naughty minimization term. It fights against our instincts, as we were taught as gospel that nonsmooth transformations wreak havoc with probability distributions.

Students at this point typically nod with deer-like expressions on their faces, assuming the professor knows some black magic that leads to a proof. In fact, he/she is probably employing the trusted Jedi mind trick, taking a tip from good ol' Fermat with his famous margin scribbles. (If you substitute "reader" for "student" and "author" for "professor", this description describes hundreds if not thousands of academic papers.)

Proof.

First, the *detailed balance criterion* of a Markov Chain with transition distribution $T(x_{k+1}|x_k)$ states that a stationary distribution $P(x)$ satisfies $T(x_{k+1}|x_k)P(x_k) = T(x_k|x_{k+1})P(x_{k+1})$. This is a *necessary* condition for a random walk to asymptotically reach a stationary distribution; additional considerations like ergodicity (that is, there is a nonzero probability of reaching any state from any other state) must be

satisfied to make the walk truly converge to P. A good professor will wave his/her hands about detailed balance. Few tread into the next steps of the proof.

Now, we must examine whether each subsequent iterate of the MH algorithm satisfies detailed balance. The argument is somewhat subtle.

The transition distribution of the MH sampling sequence is given by distribution of x_{k+1} *after each inner MH loop completes*, given the value of x_k at the beginning of the loop. Examine the following equations. Let Case A denote the case where $x_{k+1} \neq x_k$ and let Case B indicate $x_{k+1} = x_k$. A Case A transition can only be achieved with an accepted MH step, which happens with probability:

$$T(x_{k+1}|x_k) = \alpha(x_{k+1}, x_k)Q(x_{k+1}|x_k).$$

A Case B transition can be achieved with an accepted step that luckily lands back at the same point as well as a rejected step:

$$T(x_k|x_k) = \alpha(x_k, x_k)Q(x_k|x_k) + \int_{x'} Q(x'|x_k)\big(1 - \alpha(x', x_k)\big)dx'$$

The first term is the probability of luckily being accepted back on the same point, and the second term is the probability of a rejection. Don't get hung up on the complexity of the second term, as the form of the *transition distribution for Case B doesn't really matter*. Here, we have $x_{k+1} = x_k$, and so it is trivially evident that the detailed balance condition $T(x_{k+1}|x_k)P(x_k) = T(x_k|x_{k+1})P(x_{x+1})$ is satisfied in Case B.

Now let's return to Case A and expand the α in the transition probability, then apply a bit of algebra:

$$\begin{aligned} T(x_{k+1}|x_k) &= \min\left(1, \frac{P(x_{k+1})Q(x_k|x_{k+1})}{P(x_k)Q(x_{k+1}|x_k)}\right) Q(x_{k+1}|x_k) \\ &= \frac{1}{P(x_k)} \min\big(P(x_k)Q(x_{k+1}|x_k), P(x_{k+1})Q(x_k|x_{k+1})\big) \end{aligned}$$

Observe that the two terms inside the minimization are symmetric with respect to switching of x_k and x_{k+1}. Simply switching them in the equation above, we have

$$T(x_k|x_{k+1}) = \frac{1}{P(x_{k+1})} \min\big(P(x_k)Q(x_{k+1}|x_k), P(x_{k+1})Q(x_k|x_{k+1})\big)$$

So, in Case A, we have

$$\begin{aligned} T(x_{k+1}|x_k)P(x_k) &= \frac{1}{P(x_k)} \min\big(P(x_k)Q(x_{k+1}|x_k), P(x_{k+1})Q(x_k|x_{k+1})\big) \cdot P(x_k) \\ &= \frac{1}{P(x_{k+1})} \min\big(P(x_{k+1})Q(x_k|x_{k+1}), P(x_k)Q(x_{k+1}|x_k)\big) \cdot P(x_{k+1}) \\ &= T(x_k|x_{k+1})P(x_{k+1}) \end{aligned}$$

Fulfilling the detailed balance condition as desired. Since detailed balance holds for both Case A and Case B, it holds everywhere.

Derivation of MH.

Another way to think about MH is how should we design the acceptance probability function $\alpha(x', x)$ so that detailed balance is satisfied? First, it doesn't matter what α is for Case B to hold. So, let's turn our attention to detailed balance in Case A. Expanding this out, we require that α satisfy:

$$\alpha(x_{k+1}, x_k)Q(x_{k+1}|x_k)P(x_k) = T(x_{k+1}|x_k)P(x_k) = T(x_k|x_{k+1})P(x_{k+1}) \\ = \alpha(x_k, x_{k+1})Q(x_k|x_{k+1})P(x_{k+1})$$

In other words, $\frac{\alpha(x_{k+1},x_k)}{\alpha(x_k,x_{k+1})} = \frac{Q(x_k|x_{k+1})P(x_{k+1})}{Q(x_{k+1}|x_k)P(x_k)}$ with some algebraic manipulation.

For α to yield a valid probability, it must be nonnegative and no greater than 1. By inspecting what happens when $\alpha(x_{k+1}, x_k)$ is limited by 1, we see that $\alpha(x_k, x_{k+1})$ becomes not limited, and the equation holds as it should. The equation also holds in the opposite condition when the numerator is not limited and the denominator is. Hence, the MH acceptance probability satisfies this requirement.

Metropolis-Hastings is Optimal.

An interesting thought experiment would ask to design a *different* acceptance probability function. Let c be the rhs of the above equation. Then can we design a function $\beta(x, y, c)$ whose range is in [0,1] and satisfies $\beta(x, y, c) = c\beta(y, x, c)$ for all x,y,c?

Yes, this is obviously true for any constant scaling in (0,1], and other solutions may exist as well.

But can we design a function that leads to fewer rejections than the Metropolis acceptance probability? Metropolis has a nonzero probability of rejecting whenever $\alpha(x', x_k) = \frac{P(x')Q(x_k|x')}{P(x_k)Q(x'|x_k)}$ is less than 1. But if we were to use a higher probability of accepting the sample, say $\beta(x', x_k) > \alpha(x', x_k)$, then we'd have a problem: $\beta(x_k, x') = \beta(x', x_k)/\frac{Q(x_k|x_{k+1})P(x_{k+1})}{Q(x_{k+1}|x_k)P(x_k)} = \beta(x', x_k)/\alpha(x', x_k) > 1$. Hence, the Metropolis acceptance probability leads to a maximum number of accepted steps.