A Lesson in measure theory and change of variables

Peter Carbonetto
University of British Columbia
November 15, 2005

The snooker algorithm and its more general counterpart, adaptive direction sampling, were devised in order to alleviate some of the inherent problems of the Gibbs sampler [Gilks *et al.*, 1994]; namely, that it can be slow to converge even in unimodal situations. Gilks *et al.* have some nice figures illustrating this. The idea behind the snooker algorithm was to keep score of a population of samples that would then help each other move effectively in regions of high density. In other words, the algorithm maintains a series of Markov chains that interact through "snooker moves". From this brief description, it seems like a rather obvious forebear to the family of methods known as population Monte Carlo [Liang and Wong, 2001], the only novelty being that parallel Markov chains are drawn from different, but related, densities. Most commonly, these densities are related through a "temperature ladder". Even though Roberts and Gilks [1994] consider a less general setting, the snooker algorithm is equally valid in population Monte Carlo. In fact, it would be fair to say that it is a key ingredient in the success of population Monte Carlo, because it can push a chain in the direction of another chain (unlike the exchange move) while maintaining reversibility. However, it is also a subtle exposition of the beauty (and danger) of measure theory, and due to its popularity, it is worth taking the time to make sure we have a proper understanding of its inner workings.

Rather than a single chain, we maintain a set of points $\boldsymbol{x} \equiv \{x_1, x_2, \ldots, x_n\}$ representing states following the distribution of interest $\pi(x)$. The first step in the snooker algorithm uniformly chooses a point x_a from the family of n samples that will act as an *anchor*, and another x_c called the current point. Here the analogy to snooker (or pool) comes into play. The anchor point x_a is the cue ball, and the current point x_c is the "object ball" (in pool, it would be either a solid or a stripe, and in snooker it is the red ball). While Roberts and Gilks do admit that this analogy can only be taken so far, it is still helpful to see the snooker move as hitting the object ball along a line in the direction of the cue ball. The distance the ball travels is the random variable u, and it drawn from the density proportional to

$$\pi(x_c + u(x_a - x_c))|1 - u|^{d-1}, \tag{1}$$

where d is the dimension of the sample space (it's assumed that the samples are drawn from $\mathbb{R}^d$ so that we can conveniently work with Lebesgue measures[1]). Then the last step is to set the new value y_c of the c^{th} Markov chain to

$$y_c = x_c + u(x_a - x_c).$$

All this is, again, illustrated quite nicely with figures in [Gilks *et al.*, 1994].

[1] The Lebesgue measure corresponds to the length of the interval. If $\mu(\triangle x)$ is the Lebesgue measure with respecet to some interval $\triangle x$, then $\mu(\triangle x) = \triangle x$. While seemingly natural, the construction of the Lebesgue measure is not trivial [Pollard, 2002].

However, the snooker algorithm opens up a slew of questions, such as: what to do if we can't sample from the density given by (1)? and where does the term $|1-u|^{d-1}$ come from anyway? It turns out that answering the second question will help us in answering the first.

One surefire way to check the validity of a Markov chain is to check the reversibility condition (also called "detailed balance"), which requires

$$\iint K(dy\,|\,x)\,\pi(dx) = \iint K(dx\,|\,y)\,\pi(dy) \tag{2}$$

for some transition kernel K. If the transition kernel satisfies (2), then $\pi(dx)$ is the invariant density of $K(dx\,|\,y)$ [Chib and Greenberg, 1995].[2] Things become considerably more complicated with the snooker algorithm because we generate a random variable u, and then deterministically set the new value y_c according to u, so the kernel looks something like $K(du\,|\,x_c)$. x_c is now a function of the random draw u. What we need is a *change of variables*.

Recall the Change of Variables Theorem from calculus (see for example [Marsden and Tromba, 1999]), which states that

$$\iint_D f(x,y)\,dx\,dy = \iint_{D^\star} f(x(u,v),y(u,v))\left|\frac{\partial(x,y)}{\partial(u,v)}\right| du\,dv$$

assuming the transformation $T(u,v) = (x(u,v),y(u,v))$ from the set $D^\star$ to the set D is C^1 continuous. The determinant of the Jacobian is given by

$$\left|\frac{\partial(x,y)}{\partial(u,v)}\right| \equiv \left|\begin{array}{cc} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{array}\right|$$

This is the 2-d version theorem, but it extends to multivariate calculus in a similar manner.

Following the change of variables theorem and assuming Lebesgue measures from here on in, we can restate the detailed balance condition (2) for the snooker algorithm as

$$\iint K(u'\,|\,y)\,\pi(y)\,dy\,du' = \iint K(u\,|\,x)\,\pi(dx)\left|\frac{\partial(y,u')}{\partial(x,u)}\right| dx\,du, \tag{3}$$

such that $x \equiv x_c$ and $y \equiv y_c$.[3] However, a curious thing happens if we evaluate the Jacobian determinant in (3). The new state is given by the equation $y(x,u) = x + u(x_a - x)$, so the partial derivatives of y are

$$\frac{\partial y}{\partial x} = (1-u)I_d, \qquad \frac{\partial y}{\partial u} = x_a - x$$

where I_d is the $d \times d$ identity matrix (remember x, y and x_a are vectors in $\mathbb{R}^d$). The move in reverse is $x = y + v(x_a - y)$ for some step length $v \in \mathbb{R}$. The step length v in the reverse move can be determined by simple geometry: we need to move one unit to go back to the anchor from y, and in order to go way back to x we need to move a little bit more, $u' = u/(u-1)$. You can check this by substituting y into the reverse snooker move. Its

[2] There are other ways to check the invariant distribution of a Markov transition kernel, but they tend to be difficult. Time reversibility does not guarantee other important properties of a Markov chain, such as irreducibility.

[3] For a more rigorous and elegant exposition of the detailed balance equation, check out [Green, 2003].

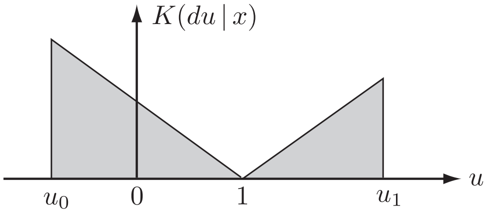

Figure 1. Density function $K(du\,|\,x)$ when $x \in \mathbb{R}^2$. u_0 and u_1 are bounds on the step size.

derivative is $\partial u'/\partial u = (1-u)^{-2}$. Everything is accounted for, so the determinant of the Jacobian is

$$\begin{vmatrix} \frac{\partial y}{\partial x} & \frac{\partial y}{\partial u} \\ \frac{\partial u'}{\partial x} & \frac{\partial u'}{\partial u} \end{vmatrix} = \begin{vmatrix} (1-u)I_d & x_a - x \\ 0\ 0\ \cdots\ 0 & (1-u)^{-2} \end{vmatrix} = \pm|1-u|^{d-2} \tag{4}$$

Since each $dx\,du$ refers to the area of a small rectangle, its transformed counterpart $|J|dx'du'$ must also be an area, hence $|J|$ must be positive. When a matrix is upper triangular, the determinant is the product of the entries on the diagonal. However, (4) is not what we wanted, since Roberts and Gilks adjust the proposal density with $|1-u|^{d-1}$! Unintuitively, the kernel

$$K(u\,|\,x, x_a) = \frac{\pi(x + u(x_a - x))|1-u|^{d-1}}{\int \pi(x + u^*(x_a - x))|1-u^*|^{d-1}du^*}. \tag{5}$$

indeed has invariant distribution $\pi(x)$. Roberts and Gilks [1994] show that the detailed balance equation (3) holds for this particular choice of kernel. While they claim the proof is straightforward, it is rather arduous, so we won't repeat it here.

Let's run a little experiment to verify these results. We use the example[4] suggested in [Gilks *et al.*, 1994]: assuming a uniform density $\pi(x)$ on the unit circle $D \equiv \{x \in \mathbb{R}^2 \text{ s.t. } \|x\| \leq 1\}$, a valid Markov chain should explore an inner circle of radius $1/2$ a quarter of the time, since (area of outer circle)/(area of inner circle) $= 4$. We compare two snooker algorithms with and without the correction term $|1-u|$. We set $n = 25$ and generate a Markov chain of length 5000. The one-dimensional density $K(u\,|\,x) = |1-u|$ looks something like the drawing in Fig. 1. See Ch. 2 in [Devroye, 1986] for hints on how to simulate this density.

After running the experiment, the uncorrected snooker algorithm gets a ratio of 2.85, far off the mark, while the corrected version gets 4.02. The density plot in Fig. 2 shows why the first algorithm fails: it spends an inordinate amount of time near the centre of the circle.

In most situations it is unrealistic to assume that we can draw samples from (5), so we explore a third option using the Metropolis-Hastings algorithm [Robert and Casella, 2004]. Via the "trans-dimensional" framework formalized by Green [2003], the Metropolis-Hastings acceptance probability under the change of variables (3) is

$$A(x,y) = \min\left\{1, \frac{\pi(y)\,q(u'\,|\,y)}{\pi(x)\,q(u\,|\,x)}\left|\frac{\partial(y,u')}{\partial(x,u)}\right|\right\},$$

[4] The explanation of this example given in [Gilks *et al.*, 1994] could be misleading. The authors seem to justify the presence of the correction term since the kernel density becomes the same as the target density. However, one could imagine adding a small bump to the region D that invalidates such an line of thinking.

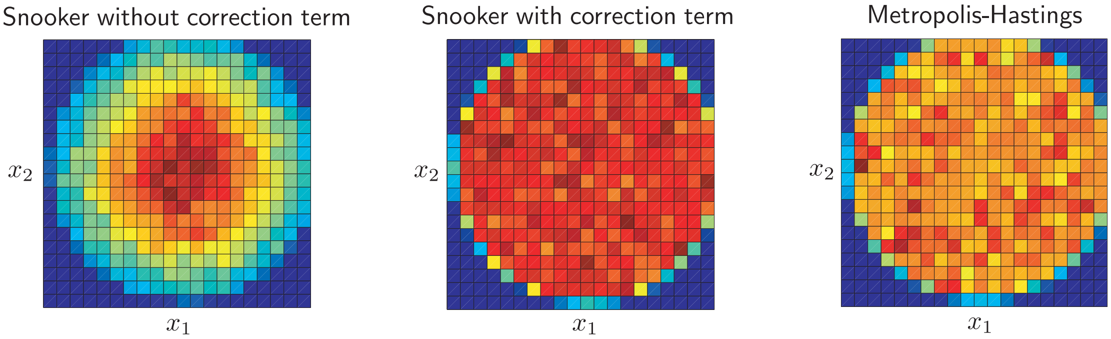

Figure 2. Density plots comparing three different MCMC algorithms: *(left)* the snooker transition kernel without the correction term, *(middle)* the snooker kernel with the correction term, *(right)* and Metropolis-Hastings. The estimated density in the first plot is not uniform.

where $q(u \mid x)$ is the probability of proposing step size u. Since we have a uniform distribution, $\pi(y)$ and $\pi(x)$ cancel. With a uniform proposal, the ratio of the proposals in the forward and backward directions is equal ratio of their normalizing constants,

$$\frac{q(u' \mid y)}{q(u \mid x)} = \frac{u_1 - u_0}{u'_1 - u'_0}.$$

Refer to Fig. 1 to see why. Here, the Jacobian is the same as usual, $|1-u|^{d-2}$. But wait! How can the Metropolis-Hastings step have a different correction term than the snooker move? And that means in our experiment, where $d = 2$, the Jacobian disappears from the acceptance probability! These are all valid observations, but Fig. 2 demonstrates that the Metropolis-Hastings algorithm we have derived here is correct. In our trial run, the estimated ratio was 4.02. Note that if we used the other correction term, we would get the wrong result.

The lesson to be learned here is that a correct application of the change of variables theorem to MCMC is not an easy task and requires some familiarity with measure theory.

References

[Chib and Greenberg, 1995] S. Chib and E. Greenberg. Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4):327–335, 1995.

[Devroye, 1986] L. Devroye. *Non-Uniform Random Variate Generation*. Springer-Verlag, 1986.

[Gilks *et al.*, 1994] W. R. Gilks, G. O. Roberts, and E. I. George. Adaptive direction sampling. *The Statistician*, 43(1):179–189, 1994.

[Green, 2003] P. J. Green. Trans-dimensional Markov chain Monte Carlo. In *Highly Structured Stochastic Systems*. Oxford University Press, 2003.

[Liang and Wong, 2001] F. Liang and W. H. Wong. Real-parameter evolutionary Monte Carlo with applications to Bayesian mixture models. *Journal of the American Statistical Association*, 96(454):653–666, 2001.

[Marsden and Tromba, 1999] J. E. Marsden and A. J. Tromba. *Vector Calculus*. W.H. Freeman and Company, 4th edition, 1999.

[Pollard, 2002] D. Pollard. *A User's Guide to Measure Theoretic Probability*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2002.

[Robert and Casella, 2004] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 2004.

[Roberts and Gilks, 1994] G. O. Roberts and W. R. Gilks. Convergence of adaptive direction sampling. *Journal of Multivariate Analysis*, 49:287–298, 1994.