CATALYTIC PERFECT SIMULATION

L.A. BREYER AND G.O. ROBERTS

Abstract. We introduce a methodology for perfect simulation using so called catalysts to modify random fields. Our methodology builds on a number of ideas previously introduced by Breyer and Roberts (1999), by Murdoch (1999), and by Wilson (1999). We illustrate our techniques by simulating two examples of Bayesian posterior distributions.

1. Introduction

With the appearance of Propp and Wilson's (1996) paper, an exciting prospect for Markov Chain Monte Carlo was unveiled: to simulate directly independent realizations from a target distribution, using existing Markov chains to drive the simulation. The basic algorithm, caled CFTP, was quickly taken to task and also extended to various settings of great generality. Various initial assumptions about the driving Markov chain, including monotonicity and uniform ergodicity were shown to be unnecessary, although they greatly simplified the implementation of these algorithms in particular cases. As a result, much effort was spent seeking out these simplifying assumptions, which has lead to a general ambivalence about the versatility of Perfect Simulation.

Our aim in this report is to combine various recent advances from several authors (Wilson, 1999, Murdoch, 1999, Breyer and Roberts, 1999) into a single, easily implementable methodology. This will be illustrated on two Bayesian inference problems: a hierarchical model example and a finite mixture example.

Our approach is fairly general, and unlike many existing methods, does not require specific knowledge of the Markov chain state space or the target distribution.

Date: June, 2000.

1991 *Mathematics Subject Classification.* Primary 60J, Secondary 60F.

Key words and phrases. Markov chains, Coupling Constructions, Gibbs samplers, Perect Simulation.

The price we pay for this generality is a loss of efficiency, due to the fact that we shall always ignore any desirable features, such as monotonicity or uniform ergodicity, which are problem specific. This loss of efficiency limits the ultimate applicability of the method, though a wide range of Bayesian applications are accessible through our methodology.

Briefly, we shall describe in this paper how to convert an existing Markov chain implementation, with prescribed stationary distribution π, so that perfect samples from π can be read off the sample path (at random time intervals).

The possibility of doing this in a general Markov chain context is due to Wilson (1999), who calls the method Read-Once CFTP and shows the relation with the PASTA principle of Applied Probability. The second ingredient upon which we base our presentation was emphasised by Murdoch (1999): to obtain a uniformly ergodic Markov chain by utilizing iterations from a merely positive recurrent chain, it suffices to insert regularly a step from an Independence Sampler with a sufficiently heavy tailed proposal distribution (usually the prior in a Bayesian posterior simulation problem). While the Independence Sampler on its own performs poorly for simulation, it is most certainly useful as a *starting point* for other, more specialized simulation steps. The third and final ingredient is a coupling construction due to Breyer and Roberts (1999), which allows a computationally straightforward approach to the implementation of Wilson's Read-Once CFTP blueprint. Examples are given in the final sections of the paper.

2. Problem setting

It is required that we should simulate random variables whose law follows a prescribed probability distribution, given by an unnormalized density $\pi(x)$ on a suitable state space E, with reference measure dx. Conventional Markov Chain Monte Carlo methodology addresses this problem by simulating a Markov chain $X_0, X_1, \ldots$ whose transition probabilities $P(x, dy) = \mathbb{P}(X_{t+1} \in dy \,|\, X_t = x)$ preserve the density π: $\int \pi(x)P(x, dy)dx = \pi(y)dy$. Starting with an arbitrary state X_0, convergence theorems guarantee that the distribution of X_t is very close to $\pi(x)/\int \pi(z)dz$, provided t is sufficiently large.

From a practical point of view, the simulation of the chain X_t is accomplished by choosing independently at each iteration a random function $F_t : E \to E$ whose defining characteristic is that $\mathbb{P}(F_t(x) \in dy) = P(x, dy)$, and setting $X_{t+1} = F_t(X_t)$.

There is considerable freedom in the choice of $P(x, dy)$ (subject to preserving the density $\pi(x)$), and more so in the choice of realizations F_t. Rather than enter a debate about the relative merits of these choices, we suppose here that the end-user has already made a decision about how to update the state. Accordingly, let us state the following assumptions

Assumption I: A source of independent random update functions $f(x)$ exists, each satisfying

$$\int \pi(x)\mathbb{P}\Big(f(x) \in dy\Big)dx = \pi(y)dy,$$

and the simulation of π should use updates of this type. The target density $\pi(x)$ is known up to a normalization constant.

Assumption II: The update functions f have a common probability density $p(x, y)$, which is known up to a normalizing constant:

$$\mathbb{P}\Big(f(x) \in dy\Big) = p(x, y)dy, \quad x \in E. \tag{1}$$

Ordinarily, simulation proceeds by updating $X_0 \mapsto f_1(X_0) \mapsto f_2 \circ f_1(X_0) \mapsto \cdots$, but the version of Perfect Simulation we propose here will identify random times when the state is exactly distributed according to π. This requires some mild modifications of the update rules, which can be performed generically in a usually straightforward manner. To support this claim, the bulk of this report consists of a selection of realistic examples.

3. Updating rules

Given that we wish to use the update functions from Assumption I, it is still necessary to perform certain modifications which allow the machinery of Perfect Simulation to operate. These modifications result in a sequence of independent compound updates, denoted $F_1, F_2, \ldots$, which are based upon the sequence $f_1, f_2, \ldots$ and can be used to define a Markov chain $X_{t+1} = F_t(X_t)$ and random times $T_1, T_2, \ldots$ whose definition depends on the F_t, such that $X_{T_k} \sim \pi$ for $k = 2, 3, \ldots$. To construct each F_t, we make use of the following ideas:

(a) Resetting the state: Let $b(x)$ denote a density on E which can be easily simulated from, and dominates π in the sense that $\pi(x)/b(x) \to 0$ as $x \to \infty$ (when the state space E is bounded, this requirement is vacuous). For a realization $B \sim b$, we define the random update function

$$R_B(x) = \begin{cases} B & \text{if } \pi(B)b(x) > \psi\pi(x)b(B) \\ x & \text{otherwise,} \end{cases} \tag{2}$$

where $\psi \sim U[0,1]$ independently. This update is the building block of the Markov chain known as the Independence Sampler.

(b) Coupling the updates $f_1, f_2, \ldots$: Given an update function $x \mapsto f(x)$ as in Assumption I, we can define a modification $\boldsymbol{\mathcal{C}}_Y(f)$ as follows: Let $q(x)$ be a probability density on E which can be easily simulated, and let Y be a realization from q. We set

$$\boldsymbol{\mathcal{C}}_Y(f)(x) = \begin{cases} Y & \text{if } p(x,Y)q(f(x)) > \xi p(x,f(x))q(Y) \\ f(x) & \text{otherwise,} \end{cases} \tag{3}$$

where $\xi \sim U[0,1]$ independently, and $p(x,y)$ comes from Assumption II. Choose a finite sequence $\mathcal{Y} = (Y_1, Y_2, \ldots, Y_\tau)$ of IID random variables from q. We can extend the above definition by writing

$$\boldsymbol{\mathcal{C}}_{\mathcal{Y}}(f) := \boldsymbol{\mathcal{C}}_{Y_1,Y_2,\ldots,Y_\tau}(f) = \boldsymbol{\mathcal{C}}_{Y_\tau} \circ \boldsymbol{\mathcal{C}}_{Y_{\tau-1}} \circ \cdots \circ \boldsymbol{\mathcal{C}}_{Y_2} \circ \boldsymbol{\mathcal{C}}_{Y_1}(f). \tag{4}$$

(c) Defining the compound updates $F_1, F_2, \ldots$: Let m be a desired number of iterations. Each update $x \mapsto F_t(x)$ is constructed independently using the formula

$$F(x) = \boldsymbol{\mathcal{C}}_{\mathcal{Y}_m}(f_m) \circ \cdots \circ \boldsymbol{\mathcal{C}}_{\mathcal{Y}_1}(f_1) \circ R_B(x), \tag{5}$$

where $f_1, \ldots, f_m$ are IID updates from the source in Assumption I, $\mathcal{Y}_1, \ldots, \mathcal{Y}_m$ are independent random vectors as described in (b), and B is an independent random variable chosen as in (a).

4. Performing the simulation

Generating samples from π proceeds as follows: first we generate the random maps $F_1, F_2, \ldots$ as in (c) of the previous section, and define a Markov chain $X_{t+1} = F_{t+1}(X_t)$, X_0 arbitrary. An important reason for defining the maps F_t

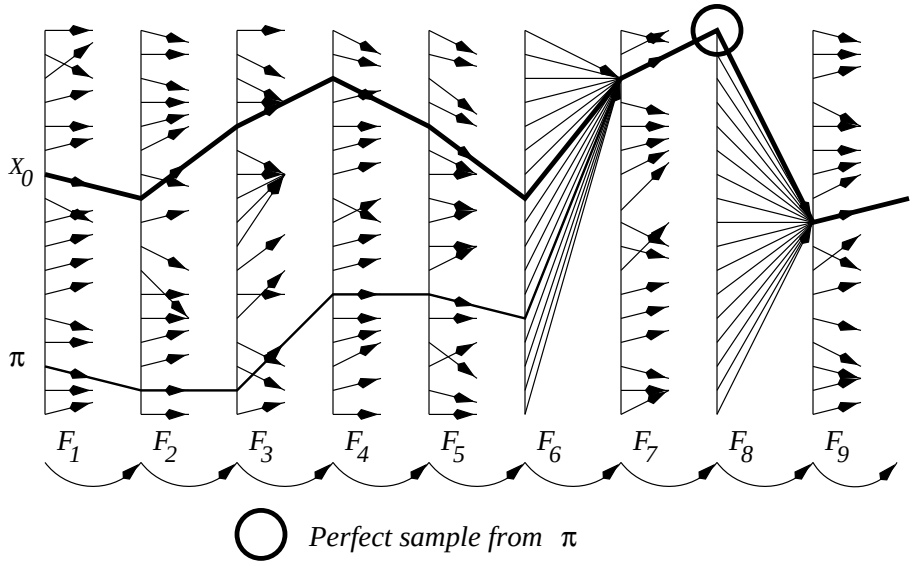

FIGURE 1. An illustration of Wilson's (1999) Read-Once CFTP construction

as we have is to allow us to test easily for the event $\{x \mapsto F_t(x) \text{ is coalescent}\}$ (a map $x \mapsto F(x)$ is coalescent if all initial points $x \in E$ are mapped onto a single final point). In particular, had we only used the given update functions $f_1, f_2, \ldots$ without modification, this event may never occur. Let us define the random variables

$$T_1 = \min\{s > 0 : x \mapsto F_{s+1}(x) \text{ is coalescent}\},$$

$$T_{k+1} = \min\{s > T_k : x \mapsto F_{s+1}(x) \text{ is coalescent}\}, k \geq 1.$$

We claim that $X_{T_2}, X_{T_3}, \ldots$ are independent samples each with distribution π (but not X_{T_1}). This is shown in the appendix. Essentially, we want to begin the simulation from a coalesced state (which is accomplished by rejection sampling, and occurs at time $T_1 + 1$), and then we must wait until just before the next coalescence (the time T_2, which is *not* a stopping time for the chain X_t).

5. Testing for coalescence

To test if a given realization of the update map $x \mapsto F_t(x)$ is coalescent, we must in principle trace all possible updates $x \mapsto F_t(x), x \in E$ and test that they all output the same state.

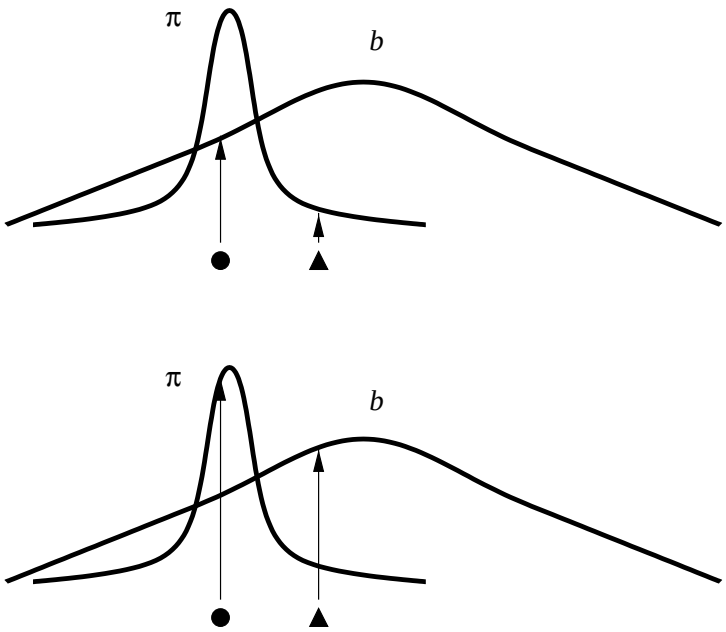

FIGURE 2. Which is more likely? The Independence Sampler update $x \mapsto R_B(x)$ in action.

In practice, a finite number of input states $x_1, \ldots, x_n$ needs to be updated and compared only, each of which is representative of a continuum of neighbouring chains which F_t maps to the same value. This is ensured by our construction of F_t, which takes care of selecting the number and location of these representative states. Thus there is no need to settle for approximate coalescence, unless the computations involved have an unacceptably high computational complexity (as might be the case for very difficult problems). We shall present two examples later. The pump example, which represents an "easy" problem, and a mixture example, which represents a class of notorioulsy difficult problems. Not surprisingly, the latter example is less satisfactory than the former.

We explain now the meaning of the elements which enter in the construction of F as given in (5).

Consider first the update $x \mapsto R_B(x)$ described in (a). Most distributions encountered in simulation are tight: for every $\epsilon > 0$, there exists a compact set K containing most of the mass, $\pi(K)/\pi(E) \geq 1 - \epsilon$. Given an arbitrary initial point $x \in E$, if we simply use the f updates repeatedly, the resulting path may take an

inordinate amount of time to reach the set K, after which the path leaves this set extremely infrequently.

To reduce wasted iterations, we could propose a global move directly into K to begin with. As the set K is not actually known, we can guess using an Independence Sampler: we update $x \mapsto R_B(x)$, where R_B is given in (a) above, and only then perform iterations based on f_t. During this first update, each possible x is either moved to equal the random variable B, or else left untouched. More precisely (see Figure 5), we compare the joint likelihood that $x \sim \pi$ and $B \sim b$ with the likelihood that $x \sim b$ and $B \sim \pi$. If the latter hypothesis is accepted, we move x to B since this realization is more typical of π. Note that, as x grows unbounded, this will always be the case since the acceptance ratio in the definition of R_B tends to infinity. The overall effect is to randomly replace the full state space E by a subset $K_B = \{x \in E : b(x)/\pi(x) \leq \psi b(B)/\pi(B)\}$. Points $x \notin K_B$ are lumped into the realization B, and the choice of K_B is unbiased, since $x \mapsto R_B(x)$ preserves the target density π.

Referring to the definition of the random maps F_t and then Figure 5, we can test for coalescence by only propagating initial points x which belong to K_B (together with the point located at B, which always belongs to K_B).

Next, we discuss the meaning of the updates $x \mapsto \boldsymbol{\mathcal{C}}_{\mathcal{Y}_k}(f_k)(x)$ defined in (b). When $k = 1$, this will also give the rule for choosing the representative points. To generate the map $x \mapsto \boldsymbol{\mathcal{C}}_Y(f)(x)$, we need both Assumptions I and II, and use the formula (4). Of crucial importance is the fact that we do not destroy the marginal distributions of f:

$$\mathbb{P}\Big(\boldsymbol{\mathcal{C}}_Y(f)(x) \in dy\Big) = p(x,y)dy = \mathbb{P}\Big(f(x) \in dy\Big).$$

Thus, even though we shall use $\boldsymbol{\mathcal{C}}_Y(f)$ to drive the simulations, convergence to stationarity is not adversely affected.

Figure 5 gives an interpretation of the effect of (4). A common value Y ("catalyst") is proposed to each $x \in E$ instead of $f(x)$, and this is always accepted by some of these. Repeating the procedure for a suitably long list $\mathcal{Y} = (Y_1, Y_2, \ldots, Y_\tau)$ ensures that all points x in any reasonable set forget the value $f(x)$ in favour of some Y_k. Here τ is taken independently of the realized sequence $Y_1, Y_2, \ldots$ and of

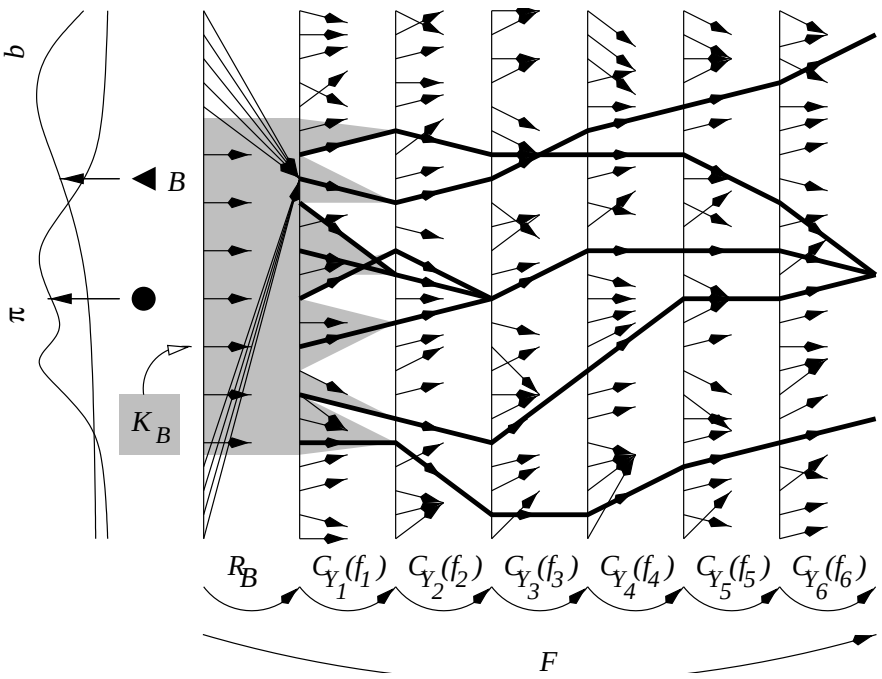

FIGURE 3. Tracking coalescence of the maps F_t using representative chains. Thin arrows represent the modified maps after the catalysts take effect, heavy lines represent the paths of the representative chains.

f, otherwise we can loose the property

$$\int \pi(x)\mathbb{P}\Big(F(x) \in dy\Big)dx = \pi(y)dy.$$

When we construct the first update $x \mapsto \boldsymbol{\mathcal{C}}_{\mathcal{Y}_1}(f_1)$ after $R_B(x)$ in (5), we must ensure that every point $x \in K_B$ is mapped onto some $Y_k, k \leq \tau$. These points $Y_1, \ldots, Y_\tau$ are then representative and can be used to trace coalescence in the map F (defined in (5)), as illustrated in Figure 5.

There exists a method (Breyer and Roberts, 1999) for generating automatically a sufficient number of points $Y_1, \ldots, Y_\tau$, but from a practical point of view, it may often be simpler to select a fixed (possibly large) number of proposals τ, and simply testing if $x \mapsto \boldsymbol{\mathcal{C}}_{\mathcal{Y}}(f)$ maps all points $x \in K_B$ to some Y_k. From a practical viewpoint, this requires us to track sets of points:

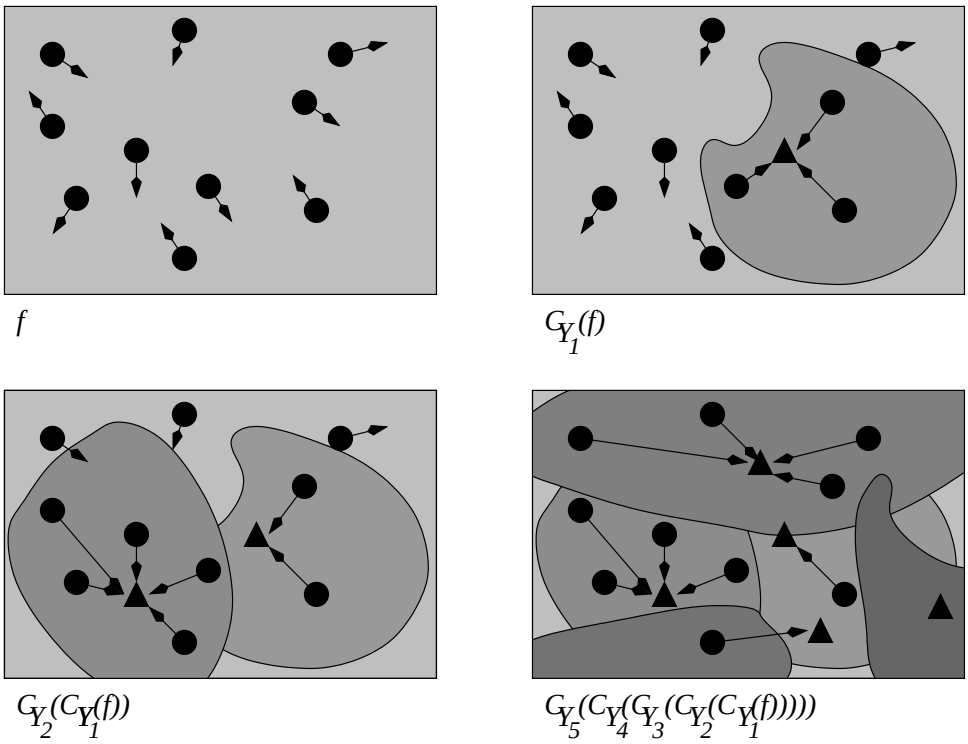

FIGURE 4. Modifications of an update map $x \mapsto f(x)$ using catalyst proposals. Arrows represent the evolution of points marked by filled in circles, triangles are the proposed catalyst values.

Definition 1. *The basin of attraction of* Y *in* (3) *is the random set*

$$Basin(Y, f, \xi) = \Big\{ x : p(x, Y)q(f(x)) > \xi p(x, f(x))q(Y) \Big\}. \tag{6}$$

If we find that not all points $x \in K_B$ are covered, we say that the whole map F doesn't coalesce.

In general, we need to ensure that the *first* update $x \mapsto \boldsymbol{\mathcal{C}}_{\mathcal{Y}_1}(f_1)$ in (5) takes all points in K_B to some Y_k. The remaining updates can have a smaller number of catalysts if desired (at least one catalyst is needed however, to guarantee that coupling between the various paths is possible).

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$	$k=9$	$k=10$
t_k	94.32	15.72	62.88	125.76	5.24	31.44	1.048	1.048	2.096	10.48
s_k	5	1	5	14	3	19	1	1	4	22

FIGURE 5. Pump dataset

6. Pump example

In this example, we describe how to simulate exactly from the posterior distribution on $\mathbb{R}_+^{11}$ given by

$$\pi(x) = \exp\left\{ (10\alpha + \gamma - 1)\log\beta - \delta\beta + \sum_{k=1}^{10}\Big((s_k + \alpha - 1)\log\lambda_k - (\beta + t_k)\lambda_k \Big)\right\},$$

where $x = (\beta, \lambda_1, \ldots, \lambda_{10})$. This model has been considered by various authors (Gelfand and Smith, 1990; Reutter and Johnson, 1995; Murdoch and Green, 1998). In the above, the constants s_k and t_k are given in Figure 6, while $\alpha = 1.802$, $\gamma = 0.01$ and $\delta = 1$.

To simulate from π, we shall use a Gibbs sampler. Here this means Assumption I holds with the typical update function given by $f : (\beta, \lambda_1, \ldots, \lambda_{10}) \mapsto (\beta', \lambda_1', \ldots, \lambda_{10}')$, where

$$\beta' \sim \pi_0(\cdot \,|\, \lambda_1, \ldots, \lambda_{10}) = \Gamma(\gamma + 10\alpha, \delta + \sum_{k=1}^{10}\lambda_k),$$

$$\lambda_k' \sim \pi_k(\cdot \,|\, \beta') = \Gamma(\alpha + s_k, \beta' + t_k), \quad k = 1, \ldots, 10.$$

Writing the density of a $\Gamma(a, b)$ random variable as $h(z) = z^{a-1}b^a \exp(-zb)/\Gamma(a)$, we get the explicit formulae

(7)

$$\pi_0(b \,|\, \lambda_1, \ldots, \lambda_{10}) = \Gamma(\gamma + 10\alpha)^{-1} b^{\gamma + 10\alpha - 1}\Big(\delta + \sum_{k=1}^{10}\lambda_k\Big)^{\gamma + 10\alpha} \exp\left(-b\Big(\delta + \sum_{k=1}^{10}\lambda_k\Big)\right)$$

$$\pi_k(l \,|\, \beta') = \Gamma(\alpha + s_k)^{-1} l^{\alpha + s_k - 1}\Big(\beta' + t_k\Big)^{\alpha + s_k} \exp\left(-l\Big(\beta' + t_k\Big)\right), \tag{8}$$

and the density for the update function f is then given by

$$p(\beta, \lambda_1, \ldots, \lambda_{10}; b, l_1, \ldots, l_{10}) = \pi_0(b \,|\, \lambda_1, \ldots, \lambda_{10}) \prod_{k=1}^{10} \pi_k(l_k \,|\, b).$$

In practical terms, the update function $x \mapsto f(x)$ is easy to implement: let $\psi_0 \sim \Gamma(\gamma+10,1)$, $\psi_1 \sim \Gamma(\alpha+s_1,1)$, $\ldots$,$\psi_{10} \sim \Gamma(\alpha+s_{10},1)$ and write $|\lambda| = \sum_{k=1}^{10}\lambda_k$. Then

$$f(\beta,\lambda_1,\ldots,\lambda_{10}) = \Big(\frac{\psi_0}{\delta+|\lambda|}, \frac{\psi_1}{\frac{\psi_0}{\delta+|\lambda|}+t_1}, \ldots, \frac{\psi_{10}}{\frac{\psi_0}{\delta+|\lambda|}+t_{10}}\Big) \tag{9}$$

represents one sweep (update) of the Gibbs sampler (we have used the fact that $X/c \sim \Gamma(a,bc)$ whenever $X \sim \Gamma(a,b)$).

We now explain how to construct the modified update $x \mapsto F(x)$ defined in (5). For the map $R_B(x)$, whose definition occusr in (2), we take $B = (B_0,\ldots,B_{10})$ such that $B_0 \sim \Gamma(\gamma,\delta)$, $B_k \sim \Gamma(\alpha,B_0)$ for $k \geq 1$. With this choice we have, in the notation of (2),

$$b(x) = \Gamma(\gamma)^{-1}\beta^{\gamma-1}\delta^{\gamma}\exp(-\beta\delta)\prod_{k=1}^{10}\Gamma(\alpha)^{-1}\lambda_k^{\alpha-1}\beta^{\alpha}\exp(-\lambda_k\beta).$$

Note that b is just the prior distribution without data s_k, t_k. A straightforward calculation gives

$$b(x)/\pi(x) = \Gamma(\gamma)^{-1}\Gamma(\alpha)^{-10}\delta^{\gamma}\exp\Big(-\sum_{k=1}^{10}(s_k\log\lambda_k - t_k\lambda_k)\Big).$$

Since this ratio is bounded, it is obvious that here

$$K_B = \Big\{x : \sum_{k=1}^{10}(s_k\log\lambda_k - t_k\lambda_k) \geq -\log\psi + \sum_{k=1}^{10}(s_k\log B_k - t_k B_k)\Big\}.$$

A convenient superset of K_B is

$$K_B \subset \Big\{(\beta,\lambda_1,\ldots,\lambda_{10}) : 0 \leq |\lambda| \leq \Big(\log\psi - \sum_{k=1}^{10}(s_k\log B_k - t_k B_k)\Big)/\max_j t_j\Big\}.$$

Next, consider the construction of the maps $\boldsymbol{\mathcal{C}}_Y(f)$ required to complete the construction of $F(x)$ in (5). The acceptance ratio in (3) simplifies if we take $Y \sim q$, where

$$q(y_0,y_1,\ldots,y_{10}) = \pi_0(y_0 \,|\, \lambda_1^*,\ldots,\lambda_{10}^*)\prod_{k=1}^{10}\pi_k(y_k \,|\, y_0),$$

and $(\lambda_1^*,\ldots,\lambda_{10}^*)$ is chosen arbitrarily. That is, our proposal is the result of one full seeep of the Gibbs sampler from some initial location $x^* = (\beta^*,\lambda_1^*,\ldots,\lambda_{10}^*)$.

Indeed, definition (3) now becomes, writing $|\lambda^*| = \sum_{k=1}^{10} \lambda_k^*$ for simplicity,

$$\mathcal{C}_Y(f)(x) = \begin{cases} Y & \text{if } \exp\Big(\beta'(|\lambda| - |\lambda^*|) - Y_0(|\lambda| - |\lambda^*|)\Big) > \xi, \\ (\beta', \lambda_1', \ldots, \lambda_{10}') & \text{otherwise.} \end{cases}$$

If we insert the expression $\beta' = \psi_0/(\delta + |\lambda|)$ derived from (9) the acceptance ratio, we immediately get the basin of attraction for Y, in conformance with Definition 1:

$$\text{Basin}(Y, f, \xi) = \Big\{x : \Big(\psi_0 - (\delta + |\lambda|)Y_0\Big)\Big(|\lambda| - |\lambda^*|\Big) > \log \xi\Big\},$$

and if we solve the implied quadratic equation, we arrive at

$$\text{Basin}(Y, f, \xi) = \left\{x : |\lambda| \in \Big[\frac{-b - \sqrt{b^2 - 4ac}}{2a}, \frac{-b + \sqrt{b^2 - 4ac}}{2a}\Big]\right\}, \tag{10}$$

where $a = Y_0$, $b = \log \xi - Y_0(|\lambda^*| - \delta)$ and $c = \delta(\log \xi - Y_0\,|\lambda^*|) - \psi_0$.

The rationale for choosing Y is therefore that, for each Gibbs sweep f, we propose the result of an independent sweep with initial configuration $x^* = (\beta^*, \lambda_1^*, \ldots, \lambda_{10}^*)$, and (4) then means that we ask every configuration x wether the old value $f(x)$ should be replaced by Y. We expect this to be the case for all x sufficiently close to x^*, and the expression (10) exactly quantifies this intuition. As we choose the variable x^* randomly over the state space K_B, we obtain a series of proposals whose basins of attraction hopefully cover all of K_B, as required in Figure 5.

The details given above are sufficient to implement the perfect simulation methodology described in Section 2, and we refer the reader to the URL
`http://www.maths.lancs.ac.uk/~robertgo/java/`
for a Java implementation.

7. Mixture example

In this example, we show how to sample from the posterior in a simple mixture problem, first discussed in Hobert et al. (1999). Let $p_1(v), \ldots, p_r(v)$ be a finite sequence of probability densities on $\mathbb{R}$ and let $\eta_1, \ldots, \eta_n$ be IID observations from the mixture density

$$v \mapsto \sum_{k=1}^{r} m_k p_k(v),$$

with unknown weights $(m_1, \ldots, m_r)$. When the prior is a uniform Dirichlet distribution $\mathcal{D}(1, 1, \ldots, 1)$ on the simplex of probability measures

$$S = \Big\{(m_1, \ldots, m_r) : m_k \geq 0 \text{ for all } k \text{ and } \sum_{k=1}^{r} m_k = 1\Big\},$$

the posterior to sample from is given by

$$\pi(m_1, \ldots, m_r) = \prod_{i=1}^{n} \Big[\sum_{k=1}^{r} m_k p_k(\eta_i)\Big].$$

Recall that the Dirichlet distribution $\mathcal{D}(\alpha_1 + 1, \ldots, \alpha_r + 1)$ has density on S given by

$$h_\alpha(m_1, \ldots, m_r) = \frac{\Gamma(r + \alpha_1 + \cdots + \alpha_r)}{\Gamma(1 + \alpha_1) \cdots \Gamma(1 + \alpha_r)} m_1^{\alpha_1} \ldots m_r^{\alpha_r}.$$

We can simulate such a random variable by generating $|\alpha| + r = \alpha_1 + \cdots + \alpha_r + r$ independent exponential variables with mean one, written $X_{11}, \ldots, X_{1(\alpha_1+1)}$, $X_{21}, \ldots, X_{2(\alpha_2+1)}$, ... ,$X_{r1}, \ldots, X_{r(\alpha_r+1)}$ and setting $m_k = \sum_j X_{kj} / \sum_{ij} X_{ij}$.

To simulate from π, we define auxiliary variables $Z_1, \ldots, Z_n \in \{1, \ldots, r\}$ which represent the mixture allocations of the data points, and use a Gibbs sampler:

$$\begin{aligned}(M_1', \ldots, M_r') &\sim \pi_0(\cdot \,|\, Z_1, \ldots, Z_n),\\ Z_i' &\sim \pi_i(\cdot \,|\, M_1', \ldots, M_r'), \quad i = 1, \ldots, n,\end{aligned}$$

where $\pi_0(\cdot \,|\, z_1, \ldots, z_n) = \mathcal{D}(N_1(z) + 1, \ldots, N_r(z) + 1)$ with $N_k(z) = \#\{s : z_s = k\}$, and

$$\pi_i(k \,|\, m_1, \ldots, m_r) = m_k p_k(\eta_i) \Big/ \sum_{j=1}^{r} m_j p_j(\eta_i), \quad k = 1, \ldots, r. \tag{11}$$

To satisfy Assumption I, let $f : (z_1, \ldots, z_n; m_1, \ldots, m_r) \to (Z_1', \ldots, Z_n'; M_1', \ldots, M_r')$, a random function depending on a fixed sequence of independent random variables $\psi_0, \ldots, \psi_n \sim U[0, 1]$ and $X_{ij} \sim \exp(1)$,with $i = 1, \ldots, r$, $j = 1, \ldots, n$, given by

$$M_k'(z) = \sum_{j=1}^{N_k(z)} X_{kj} \Big/ \sum_{i=1}^{r} \sum_{j=1}^{N_i(z)} X_{ij}, \quad k = 1, \ldots, r,$$

$$Z_i'(M') = \min\Big\{k : \sum_{j=1}^{k} M_j' p_j(\eta_i) \geq \psi_i \sum_{j=1}^{r} M_j' p_j(\eta_i)\Big\}, \quad i = 1, \ldots, n.$$

The corresponding density is

$$p(x, x') = \pi_0(M_1', \ldots, M_r' \,|\, z_1, \ldots, z_n) \prod_{i=1}^{n} \pi_i(Z_i' \,|\, M_1', \ldots, M_r'),$$

where $x = (z_1, \ldots, z_n; m_1, \ldots, m_r)$ and $x' = (Z'_1, \ldots, Z'_n; M'_1, \ldots, M'_r)$. Here, the state space $E = \{1, \ldots, r\}^{\otimes n} \times S$ is already compact, so there is no need to introduce an update $x \mapsto R_B(x)$ in the definition of $x \mapsto F(x)$ in (5). Instead, we concentrate on the definition of $\boldsymbol{\mathcal{C}}_Y(f)$ in (4). More precisely, we take as proposal

$$Y = (Y_1, \ldots, Y_n; U_1, \ldots, U_r) = f(z_1^*, \ldots, z_n^*; m_1^*, \ldots, m_r^*) \sim p(x^*, \cdot),$$

where f is constructed using independent random variables $\Psi_0, \ldots, \Psi_n \sim U[0,1]$ and $\chi_{ij} \sim \exp(1)$, in which case the acceptance ratio in (3) simplifies immediately to

$$\begin{aligned}\frac{\pi_0(U_1, \ldots, U_r \,|\, z_1, \ldots, z_n)\pi_0(M_1, \ldots, M_r \,|\, z_1^*, \ldots, z_n^*)}{\pi_0(M_1, \ldots, M_r \,|\, z_1, \ldots, z_n)\pi_0(U_1, \ldots, U_r \,|\, z_1^*, \ldots, z_n^*)} & \\ = \big(\frac{U_1}{M_1}\big)^{N_1(z) - N_1(z^*)} \cdots \big(\frac{U_r}{M_r}\big)^{N_r(z) - N_r(z^*)} & \\ = \exp \sum_{k=1}^{r} (N_k(z) - N_k(z^*)) \log(U_k / M_k). &\end{aligned}$$

This may be inserted into (6) to get the basin of attraction of Y,

$$\begin{aligned}\text{(12)} \quad \text{Basin}(Y, f, \xi) = \Big\{ (z_1, \ldots, z_n) : & \\ \sum_{k=1}^{r} \big(N_k(z) - N_k(z^*)\big) \log[U_k(z^*)/M_k(z)] > \log \xi \Big\}. &\end{aligned}$$

Note that this set always contains the configuration $x^* = (z_1^*, \ldots, z_n^*)$, which we think of as the "center" of the set.

Perfect simulation for this problem should now proceed as follows: beginning with in principle all possible allocations $(z_1, \ldots, z_n)$, we generate independent proposals $Y_1, \ldots, Y_\tau$ followed by a test to see if $\bigcup_{k=1}^{\tau} \text{Basin}(Y_k, f, \xi_k)$ covers all those allocations. If so, we need only track the paths of chains starting in $Y_1, \ldots, Y_\tau$.

In practice, it is somewhat time consuming to perform this test exactly and we opt instead for an approximate procedure which consists in sampling randomly from all configurations $(N_1, \ldots, N_r : \sum_{j=1}^{r} N_j = n)$ to see if each belongs to some basin $\text{Basin}(Y, f, \xi)$. While not exact, this procedure does give a degree of confidence in whether full coverage did or did not occur, and can be justified on the basis that neighbouring configurations tend to belong to the same basin. Methods for making this final step move exact are currently being investigated.

An implementation of this algorithm can be found at the URL
`http://www.maths.lancs.ac.uk/~robertgo/java/`

8. Metropolis-Hastings chains

In both of the two examples given above, the simulated Markov chains are Gibbs samplers. This has the advantage that transition probabilities for a full sweep are absolutely continuous, which allows the catalytic coupling to be implemented in a straightforward fashion.

Besides Gibbs samplers, the other class of Markov chains used for Monte Carlo simulations are Metropolis-Hastings chains. These have a slightly more complex transition rule due to the extra accept/reject mechanism. However, the transition kernel can again be written down explicitly which allows for three distinct coupling strategies.

Recall that the general Metropolis-Hastings transition kernel is in the form (cf. Roberts and Smith, 1993)

$$P(x, dy) = q(x, y)\left\{1 \wedge \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right\}\mu(dy) + r(x)\delta_x(dy), \tag{13}$$

where $q(x, y)$ is the proposal kernel density, $\pi(y)$ is the density of the stationary distribution, both with respect to the reference measure $\mu(dy)$, and $r(x)$ is chosen so that $P(x, E) = 1$.

The formula (13) cannot be directly used in the expressions (3), since the function $r(x)$ is not usually computable. Catalytic coupling can still be used in any one of the following ways:

- The simplest method consists in coupling the updates *before* the Metropolis accept/reject mechanism is applied. Here we exploit the fact that the Metropolized update rule has the special form

$$f(x) = \begin{cases} g(x, \psi) & \text{if } \pi(g(x, \psi))q(g(x, \psi), x) > \chi\pi(x)q(x, g(x, \psi)) \\ x & \text{otherwise,} \end{cases}$$

  where $\chi \sim U[0, 1]$ and ψ is a random variable such that $g(x, \psi) \sim q(x, \cdot)$. Introducing at the proposal stage independent catalysts $\mathcal{Y} = (Y_1, \ldots, Y_r)$,

each with law $h(\cdot)$, we get a random map $\boldsymbol{\mathcal{C}}_{\mathcal{Y}}(g) = \boldsymbol{\mathcal{C}}_{Y_r} \circ \cdots \circ \boldsymbol{\mathcal{C}}_{Y_1}(g)$, where e.g.

$$\boldsymbol{\mathcal{C}}_{Y_1}(g)(x) = \begin{cases} Y_1 & \text{if } q(x, Y_1)h(g(x,\psi)) > \xi q(x, g(x,\psi))h(Y_1) \\ g(x,\psi) & \text{otherwise.} \end{cases}$$

The full Metropolis update f is now modified to become

$$\boldsymbol{\mathcal{C}}_{\mathcal{Y}}(f)(x) = \begin{cases} \boldsymbol{\mathcal{C}}_{\mathcal{Y}}(g)(x) & \text{if } \pi(\boldsymbol{\mathcal{C}}_{\mathcal{Y}}(g)(x))q(\boldsymbol{\mathcal{C}}_{\mathcal{Y}}(g)(x), x) > \xi\pi(x)q(x, \boldsymbol{\mathcal{C}}_{\mathcal{Y}}(g)(x)) \\ x & \text{otherwise.} \end{cases}$$

- A second approach consists in coupling Metropolis-Hastings chains *after* the accept/reject step. This is based upon the nature of the kernel $P(x, dy)$ (equation (13)) as a mixture. While with probability $r(x)$ we have $f(x) = x$, the alternative consists in choosing $f(x) \sim h(x, \cdot)d\mu$, where $h(x, y) \propto \pi(x)q(x, y) \wedge \pi(y)q(y, x)$. This alternative field can be coulpled via any catalyst proposal with distribution $k(x)\mu(dx)$ in the way described previously. This gives the formula

$$\boldsymbol{\mathcal{C}}_{\mathcal{Y}}(f)(x) = \begin{cases} x & \text{if } f(x) = x, \\ Z \sim k(\cdot) & \text{otherwise, if } \frac{k(Z)h(x,f(x))}{k(f(x))h(x,Z)} > \chi \sim U[0,1] \\ f(x) & \text{otherwise.} \end{cases}$$

We leave to the reader the proof that this transformation preserves the transition kernel $P(x, dy)$.

- To explain the third method for introducing catalysts as a way of coupling an existing update function $f(x)$ satisfying Assumption I, and whose kernel is of type (13), we begin by viewing a Metropolis chain X_t as a bivariate process $(X_t, X_{t+1}) \mapsto (X_{t+1}, X_{t+2}) = (X_{t+1}, f(X_{t+1})$ consisting of two consecutive configurations. When viewed in this way, it becomes possible to write down

explicitly computable transition probabilities. More precisely, define

$$p(x_1,x_2,x_3) = \begin{cases} \big(1\wedge\frac{\pi(x_2)q(x_2,x_1)}{\pi(x_1)q(x_1,x_2)}\big)\big(1\wedge\frac{\pi(x_3)q(x_3,x_2)}{\pi(x_2)q(x_2,x_3)}\big) & \text{if } x_1\neq x_2 \text{ and } x_2\neq x_3,\\ \big(1\wedge\frac{\pi(x_2)q(x_2,x_1)}{\pi(x_1)q(x_1,x_2)}\big)\big(1-\frac{\pi(x_3)q(x_3,x_2)}{\pi(x_2)q(x_2,x_3)}\big)_+ & \text{if } x_1\neq x_2 \text{ and } x_2= x_3,\\ \big(1-\frac{\pi(x_2)q(x_2,x_1)}{\pi(x_1)q(x_1,x_2)}\big)_+\big(1\wedge\frac{\pi(x_3)q(x_3,x_2)}{\pi(x_2)q(x_2,x_3)}\big) & \text{if } x_1= x_2 \text{ and } x_2\neq x_3,\\ \big(1-\frac{\pi(x_2)q(x_2,x_1)}{\pi(x_1)q(x_1,x_2)}\big)_+\big(1-\frac{\pi(x_3)q(x_3,x_2)}{\pi(x_2)q(x_2,x_3)}\big)_+ & \text{if } x_1= x_2 \text{ and } x_2= x_3, \end{cases}$$

then the chain (X_t, X_{t+1}) has the transition kernel $P(x_1,x_2;dx_3,dx_4) = p(x_1,x_2,x_4)\delta_{x_2}(dx_3)\mu(dx_4)$, which compares favourably with (13). Coupling this bivariate update requires a catalyst proposal in the form (x_2, Y) where $Y\sim b(\cdot)d\mu$ for some conveniently chosen density b, but the formula is identical in form to (4), viz.

$$\boldsymbol{\mathcal{C}}_Y(f)(x_1,x_2) = \begin{cases} (x_2,Y) & \text{if } p(x_1,x_2,Y)b(f(x_2)) > \xi p(x_1,x_2,f(x_2))b(Y)\\ (x_2,f(x_2)) & \text{otherwise.}\end{cases}$$

Unlike the first coupling method described earlier, here we do not decompose the original Metropolis-Hastings update $f(x)$, which allows the coupling method to wrap around existing implementations with minimal impact. Calculationally, this method appears to be similar in complexity to the first method however.

9. Discussion

We have introduced an apparently general framework for perfect simulation. It can be applied to virtually any algorithm since all we need to know is the one-step transition density of the Markov chain, and even then we only need know it up to a normalisation constant. As such, we can apply it to Gibbs samplers where the conditionals need to be simulated by rejection sampling, and rather general Hastings algorithms as described in Section 8.

The limitation in our approach comes from the need to monitor basins of attraction in order to detect coalescence. Our mixture example illustrates that even when these basins become impossible to keep track of exactly, there are reasonable and

easy to implement approximations, of arbitrary accuracy, which can be constructed along the lines of our methodology.

We also take this opportunity to remind the reader that the choice of Markov chain (ordinary Gibbs sampler here) used to drive the simulation is crucial, since the coalescence times depend essentially on transition probabilities. As always, choosing algorithms for MCMC calculations is an art, not a science. We have shown in this paper that it is possible to design straightforward coupling methods for Markov chains commonly used in MCMC, with the ultimate aim to read off IID samples from the target distribution.

Acknowledgement. This research was supported by the European Union TMR network ERB-FMRX-CT96-0095 on "Spatial and Computational Statistics".

10. Appendix

In this section, we give an algebraic proof of the validity of Read-Once CFTP. A different proof, emphasising the connection with classical CFTP, was given by Wilson (1998).

Let $P(x, dy) = \mathbb{P}\big(F(x) \in dy\big)$, where F is a composite map as defined in Section 3, (c). By assumption, we have

$$\mathbb{P}\big(F \text{ is coalescent}\big) = \epsilon > 0,$$

say. We define further $\mu(dy) = \mathbb{P}\big(F(x_0) \in dy \,|\, F \text{ is coalescent}\big)$, which is independent of $x_0 \in E$. Consequently, there is a unique kernel Q such that

$$P(x, dy) = (1 - \epsilon)Q(x, dy) + \epsilon\mu(dy). \tag{14}$$

The main result is the following identity, written in the usual algebraic notation for kernels and measures.

Theorem 2. *If $\pi P = \pi$ and (14) holds, then*

$$\pi = \epsilon \sum_{s=0}^{\infty} (1 - \epsilon)^s \mu Q^s. \tag{15}$$

Proof. Using stationarity $\pi P = \pi$, we have

$$\begin{aligned}(1-\epsilon)^k \pi Q^k &= (1-\epsilon)^{k-1}\pi(P-\epsilon\mu)Q^{k-1} \\ &= (1-\epsilon)^{k-1}\pi Q^{k-1} - \epsilon(1-\epsilon)^{k-1}\mu Q^{k-1} \\ &= \cdots = \\ &= \pi - \epsilon\sum_{s=1}^{k}(1-\epsilon)^{k-s}\mu Q^{k-s}.\end{aligned}$$

As identities between positive kernels, these are true when applied to any bounded test function $f : E \to \mathbb{R}$. Changing variables $k - s \to s$ gives

$$(1-\epsilon)^k\langle \pi Q^k, f\rangle = \langle \pi, f\rangle - \epsilon\sum_{s=0}^{k-1}(1-\epsilon)^s\langle \mu Q^s, f\rangle.$$

Since also $\big|\langle \pi Q^k, f\rangle\big| \le \|f\|$, it is now clear that we obtain (15) by letting $k \to \infty$. □

The Read-Once CFTP method described in Section 4 can now be seen as a direct application of the identity (15).

More precisely, write $T_2 = T_1 + S_1 \circ \theta_{T_1}$, where θ_t is the usual time shift operator on sample paths. We have

$$\mathbb{P}(S_1 = s \,|\, X_0 \sim \mu) = \mathbb{P}\big(F_k \text{ does not coalesce for } k \le s, F_{s+1} \text{ coalesces}\big) = \epsilon(1-\epsilon)^s$$

and consequently

$$\begin{aligned}\mathbb{E}[f(X_{T_2})] &= \mathbb{E}[f(X_{S_1}) \,|\, X_0 \sim \mu] \\ &= \sum_{s=0}^{\infty}\mathbb{E}[f(X_s) \,|\, S_1 = s, X_0 \sim \mu]\mathbb{P}(S_1 = s \,|\, X_0 \sim \mu) \\ &= \sum_{s=0}^{\infty}\epsilon(1-\epsilon)^s\langle \mu Q^s, f\rangle = \langle \pi, f\rangle.\end{aligned}$$

The claimed independence between the variables X_{T_k} and $X_{T_{k+1}}$ is a direct consequence of the regeneration event (coalescence) which occurs at time $T_k + 1$.

References

[1] Breyer, L.A. and Roberts, G.O. (1999) *A new coupling construction for random fields.*

[2] Gelfand, A.E. and Smith, A.F.M. (1990) *Sampling-based approaches to calculating marginal densities.* Journal of the American Statistical Association, 85:398-409.

[3] Hobert, J.P., Robert, C.P. and Titterington, D.M. (1999) *On perfect simulation for some mixtures of distributions.*

[4] Meyn, S.P. and Tweedie, R.L. (1993) *Markov Chains and Stochastic Stability.*

[5] Wilson, D.B. (1999) *How to couple from the past using a read-once source of randomness.*

[6] Murdoch, D. (1999) *Exact sampling for Bayesian inference: unbounded state spaces.*

[7] Reutter, A. and Johnson, V. (1995). General strategies for assessing convergence of MCMC algorithms using coupled sample paths.

[8] Murdoch, D.J. and Green, P.J. (1998) *Exact Sampling from a Continuous State Space.*

Department of Mathematics, Lancaster University, UK

E-mail address: l.breyer@lancaster.ac.uk

Department of Mathematics, Lancaster University, UK

E-mail address: g.o.roberts@lancaster.ac.uk