BAYESIAN STATISTICS 6, pp. 301–321
J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (Eds.)

Exact Sampling for Bayesian Inference: Towards General Purpose Algorithms

PETER J. GREEN and DUNCAN J. MURDOCH
University of Bristol, UK, and Queen's University, Canada

SUMMARY

There are now methods for organising a Markov chain Monte Carlo simulation so that it can be guaranteed that the state of the process at a given time is *exactly* drawn from the target distribution. The question of assessing convergence totally vanishes. Such methods are known as exact or perfect sampling. The approach that has received most attention uses the protocol of *coupling from the past* devised by Propp and Wilson (*Random Structures and Algorithms*, 1996), in which multiple dependent paths of the chain are run from different initial states at a sequence of initial times going backwards into the past, until they satisfy the condition of coalescence by time 0. When this is achieved the state at time 0 is distributed according to the required target. This process must be implemented very carefully to assure its validity (including appropriate re-use of random number streams), and also requires use of various tricks to enable us to follow infinitely many sample paths with a finite amount of work.

With the ultimate objective of Bayesian MCMC with guaranteed convergence, the purpose of this paper is to describe recent efforts to construct exact sampling methods for continuous-state Markov chains. We review existing methods based on gamma-coupling and rejection sampling (Murdoch and Green, *Scandinavian Journal of Statistics*, 1998), that are quite straightforward to understand, but require a closed form for the transition kernel and entail cumbersome algebraic manipulation. We then introduce two new methods based on random walk Metropolis, that offer the prospect of more automatic use, not least because the difficult, continuous, part of the transition mechanism can be coupled in a generic way, using a proposal distribution of convenience.

One of the methods is based on a neat decomposition of any unimodal (multivariate) symmetric density into pieces that may be re-assembled to construct any translated copy of itself: that allows coupling of a continuum of Metropolis proposals to a finite set, at least for a compact state space. We discuss methods for economically coupling the subsequent accept/reject decisions.

Our second new method deals with unbounded state spaces, using a trick due to W. S. Kendall of running a coupled dominating process in parallel with the sample paths of interest. The random subset of the state space below the dominating path is compact, allowing efficient coupling and coalescence.

We look towards the possibility that application of such methods could become sufficiently convenient that they could become the basis for routine Bayesian computation in the foreseeable future.

Keywords: BISECTION COUPLER; COUPLING FROM THE PAST; EXACT SAMPLING; GAMMA COUPLING; PERFECT SAMPLING; REJECTION COUPLER.

1. INTRODUCTION

Markov chain Monte Carlo has made something of an impact on Bayesian analysis. Many computations that we need to do but which were essentially impossible using available numerical methods can now be done. In consequence, Bayesian statistics has become mainstream, not only for routine analysis, but a method of choice for challenging inferential problems with large-scale datasets arising in modern science and technology. In the process of this development, sample-based computation, by Markov chain simulation or otherwise, has become respectable: we are

gradually realising the vast opportunities for subtle extraction of information from posterior distributions that are now available. Sample-based computation having become acceptable as a routine approach to Bayesian computation, two aspects of Markov chain sampling remain problematical: the issues of convergence — when can I assume that my Markov chain is sampling in equilibrium? — and dependence — how precise are estimates of probabilities and expectations obtained from the simulation? Arguably, the second of these is relatively easily handled, by one or other estimator of Monte Carlo variance, once convergence is assured. There are now methods for organising a MCMC simulation so that it can be guaranteed that the state of the process at a given time is *exactly* drawn from the target distribution. The question of assessing convergence totally vanishes. Such methods are known as exact or perfect sampling.

The purpose of this paper is to describe recent efforts to construct exact sampling methods for continuous-state Markov chains, and to look towards the possibility that application of such methods could become sufficiently convenient that they could become the basis for routine Bayesian computation in the foreseeable future. We remain open-minded about this possibility, and certainly cannot yet demonstrate its probability, but we are optimistic.

1.1. *Coupling*

Exact sampling is a recent discovery. Two papers first circulated in 1995, by Propp and Wilson and by Fill, set out two different schemes, both based on ideas of coupling, for adapting MCMC simulation to the task of producing a guaranteed target draw. Propp and Wilson's proposal of *coupling from the past* is easier to implement and has received more attention, and we shall concentrate on it here.

Coupling has been a valuable tool in probability theory for some time, but its use in statistical simulation has been minimal. Coupling refers to the idea of constructing the random variables of interest on a common probability space, so that questions can be answered by probabilistic rather than analytic methods. There are, for example, particularly elegant proofs of Poisson approximation theorems, or of the ergodic theorem for Markov chains, based on coupling arguments (see Lindvall, 1992, for an enthusiastic course on this subject). In simulation, the 'common probability space' becomes concrete as the multiple use of sequences of random numbers in order to introduce dependence between sample paths. In making this correspondence, all we are doing is being literal about what random variables and stochastic processes are, and thinking of our simulation programs as defining appropriately measurable functions: it is a natural perspective for a probabilist, but perhaps novel for the practising statistician.

In Markov chain simulation, coupling allows us to create sample paths from many — perhaps infinitely many — initial states simultaneously. By reusing the same random numbers, these paths are made dependent, and by appropriate design can be made to satisfy desired conditions, such as having high probability that their paths merge, or *coalesce*. Of course, this constructed dependence does not infringe the Markov property, which only requires that we use different random numbers at different times within one path, and is silent about the same times on different paths.

The essence of the game will already be evident — coupling seems to offer the prospects of constructing multiple paths that ultimately coalesce, and hence 'forget' their initial conditions. Forgetting initial conditions is exactly what we want our MCMC simulations to do. Various lines of enquiry prior to Propp and Wilson were concerned with using these ideas to diagnose convergence (for example, Johnson, 1996, Asmussen *et al.* , 1992).

1.2. *Coupling from the Past*

The leap made by Propp and Wilson was to note that coupling could be used more explicitly still, to guarantee rather than assess convergence. Informally, the idea is to conduct the simulation

from all possible states at a random (but finite) time in the past, in such a way that all paths have coalesced into a single track by time 0, and the state occupied at time 0 is a draw from the target.

The structure of algorithms based on this idea is as follows. We consider runs of length M ending at time 0, for some arbitrary value of $M > 0$, starting at every possible state of the chain at time $t = -M$. Using a sequence of random numbers, we follow the paths from each of these states forward in time; occasionally paths will coalesce, and eventually all of the paths will have coalesced into one. If this has happened when we reach time 0, then we are done. If there are still multiple possible states, we choose a larger value of M and start again, using the *same realisation* of the sequence of random numbers.

To formalise this process, to see why it works, and to describe algorithms compactly, we introduce the *stochastic recursive sequence* representation of a Markov chain. Any Markov chain $\{X_t, t = \ldots, -1, 0, 1, 2, \ldots\}$ with state space χ can be represented in the way it is simulated:

$$X_{t+1} = \phi(X_t, U_{t+1}) \tag{3.4}$$

where $\{U_t, t = \ldots, -1, 0, 1, 2, \ldots\}$ is a sequence of independent random quantities from some fixed distribution, and ϕ is an appropriate deterministic *structure function*. For example, the simultaneous-update random walk Metropolis algorithm for an arbitrary target density π can be expressed via $\phi(X, U) = X + U^{(2)}$ if $U^{(1)} < \pi(X + U^{(2)})/\pi(X)$ or X otherwise, where $U^{(1)}$ is a $U(0, 1)$ random variable and $U^{(2)}$ is a random increment of the same dimension as X with a distribution symmetric about 0.

In terms of the structure function, coupling from the past (CFTP) can be expressed in pseudo-code as follows:

```
CFTP(M):
  t ← −M
  B_t ← χ
  while t < 0
    t ← t + 1
    B_t ← φ(B_{t−1}, U_t)
  if #B_0 = 1 then
    return(B_0)
  else
    CFTP(2M)
```

and is initiated by starting with an arbitrary positive integer M, often 1. We use B_t to denote the set of states occupied at time t by the paths under consideration. Coupling from the past is based on the observation that if the chain was (conceptually) run from all initial states in χ at time $-\infty$ ($B_{-\infty} = \chi$), and these paths coalesced to a singleton by time 0 ($B_0 = \{X_0\}$), then X_0 would be drawn from the stationary distribution π. This is Theorem 2 of Propp and Wilson (1996). For certain structure functions, if this occurred, then there would be a *last* time $-T$ before 0 such that the paths from $B_{-T} = \chi$ coalesced to a singleton. The CFTP algorithm listed above is a search for T. We do not need to find it exactly: since $B_t = \chi$ for all $t < -T$, it is sufficient to check whether $T \leq M$, for $M = 1, 2, 4, \ldots$.

The CFTP algorithm is only guaranteed to terminate when using such structure functions. Foss and Tweedie (1998) showed that this requires uniform ergodicity of the Markov chain. However, variations on CFTP such as those described in Section 4 allow it to be applied in greater generality.

1.3. *Current Implementations of Exact Sampling*

As in the early days of ordinary MCMC, most current applications of CFTP have been to exact sampling of large discrete physical systems and spatial processes. Many of these exploit a *monotonicity* property, or its opposite, anti-monotonicity. A monotonic MCMC sampler is one implemented by a structure function ϕ that is order-preserving with respect to some partial order $\leq$ on the state space: $x \leq y \Rightarrow \phi(x,u) \leq \phi(y,u)$ for all u, where $\leq$ has the property that there are unique minimal and maximal elements $\widehat{0}$ and $\widehat{1}$. In the case of anti-monotonicity the ranking is reversed: $\phi(x,u) \geq \phi(y,u)$. In either case, the progress forward of paths from $B_{-M} = \chi$ to $t = 0$ can be followed by tracing only the paths from $\widehat{0}$ and $\widehat{1}$.

Papers exploiting CFTP implemented in this way include Propp and Wilson's original (1996) paper, with applications to random dimer and cluster models. Häggstrom *et al.* (1999), and Häggstrom and Nelander (1997a,b) treat area-interaction point processes, Ising models with negative interaction, the random cluster model, and random q-colourings.

Kendall (1998) and Häggstrom *et al.* (1999) go beyond straight CFTP to perform exact sampling for a chain that is not uniformly ergodic; they use a coupled dominating process, and implement the method for various spatial processes. The idea of the dominating chain is borrowed for the methods we introduce in Section 4 of this paper.

Propp and Wilson (1998) derive 'generic' methods for exact sampling on Markov chains, that do not exploit any special features such as monotonicity to facilitate coalescence. Implementations are to random spanning trees and other graphical algorithms.

Bayesian uses for exact sampling have been so far rare. Fismen (1997), in a Master's thesis supervised by Håvard Rue, applies both monotone CFTP and Fill's algorithm to idealised binary image analysis problems based on Ising model priors, and simple forms of degradation for which the likelihood does not destroy the monotonicity property. Murdoch and Green (1998) apply their methods (that do not use monotonicity, and are reviewed briefly in the next section) to a hierarchical gamma model and the well-known 'pump-reliability' data, and Fismen discusses how to adapt their methods to a hierarchical model for binomial data. Møller (1999) also addresses the pump data, using his approach to conditionally-specified models exploiting (anti-) monotonicity, that gives exact sampling up to an arbitrarily small numerical tolerance.

New work in the area of exact sampling is appearing almost weekly: the interested reader is directed to David Wilson's online annotated bibliography

`http://dimacs.rutgers.edu/~dbwilson/exact`

for the latest news.

2. EXACT SAMPLING IN CONTINUOUS STATE SPACES

When we started working on exact sampling, we were unsure that exact CFTP would ever be possible in continuous state spaces, and thought that some approximation to tolerance ϵ would be the best that could be achieved. Then we discovered several prototypical constructions that did allow it exactly, at the price generally of some tedious and cumbersome algebra; these were introduced and discussed in full in Murdoch and Green (1998). In Sections 2.1 and 2.2 below we briefly describe the *multigamma* and *rejection* couplers. In practice, in implementing CFTP, we have found it useful often to carry forward a larger but more tractable set of states than $\phi(B_{t-1}, U_t)$ — this will in general delay coalescence, but save computer time and storage. So we usually code the update as $B_t \leftarrow \Phi(B_{t-1}, U_t)$, where $\Phi(B,U)$ is any set containing $\{\phi(x,U), x \in B\}$. For example, we might simply take $\Phi(B,U) = \chi$ if B is infinite, or exactly $\{\phi(x,U), x \in B\}$ otherwise.

2.1. *Multigamma Coupling*

Our first method applies when we know explicitly the update kernel density $f(\cdot \mid x)$ of the Markov chain, the density of X_{t+1} given that $X_t = x$. Furthermore, we suppose that f is bounded below uniformly in x: $f(y \mid x) \geq r(y) \forall x, y \in \chi$ for some nonnegative function $r(\cdot)$, for which $\rho = \int r(y)dy > 0$. Then we can express every $f(\cdot \mid x)$ as a fixed-weight mixture of two distributions, one of which does not depend on x. In the one dimensional case, for example, we can write

$$p(X_{t+1} \leq y \mid X_t = x) = \rho R(y) + (1 - \rho)Q(y \mid x)$$

where

$$R(y) = \rho^{-1} \int_{-\infty}^{y} r(v)dv \quad \text{and} \quad Q(y \mid x) = (1 - \rho)^{-1} \int_{-\infty}^{y} [f(v \mid x) - r(v)]dv.$$

Assuming the inverses of R and $Q(\cdot \mid x)$ are available, the update function to be used by the algorithm is

$$\phi(x, U) = \begin{cases} R^{-1}(U^{(2)}) & \text{if } U^{(1)} < \rho \\ Q^{-1}(U^{(2)} \mid x) & \text{otherwise}, \end{cases}$$

using a pair $U = (U^{(1)}, U^{(2)})$ of Uniform random numbers. Then it is easy to see that in the `CFTP` algorithm we can use

$$\Phi(B_{t-1}, U_t) = \begin{cases} R^{-1}(U^{(2)}) & \text{if } U_t^{(1)} < \rho \\ Q^{-1}(U^{(2)} \mid X_{t-1}) & \text{if } U_t^{(1)} \geq \rho \text{ and } B_{t-1} = \{X_{t-1}\} \\ \chi & \text{otherwise}, \end{cases}$$

and the algorithm evidently terminates. In fact, T has a geometric distribution with mean $1/\rho$.

While it is often difficult or impossible to find a uniform non-zero lower bound on the update densities, quite commonly the state space can be partitioned into a finite collection of disjoint cells $\mathcal{A} = \{A_i, i = 1, \ldots, m\}$, and separate lower bounds can be placed on the densities within each cell. That is,

$$f(y \mid x) \geq r_i(y) \quad \forall x \in A_i, \quad \forall y \in \chi.$$

We further assume that $\int r_i(y)dy = \rho > 0$ is constant over all cells; this assumption can be relaxed. Then a partitioned version of the multigamma coupler is available. It uses

$$\Phi(B_{t-1}, U_t) = \begin{cases} \bigcup_{i: A_i \cap B_{t-1} \neq \emptyset} \{R_i^{-1}(U_t^{(2)})\} & \text{if } U_t^{(1)} < \rho \\ \bigcup_i \{Q_i^{-1}(U_t^{(2)} \mid x) : x \in A_i \cap B_{t-1}\} & \text{if } U_t^{(1)} \geq \rho \text{ and } \#B_{t-1} < \infty \\ \chi & \text{otherwise}. \end{cases}$$

2.2. *Rejection Coupling*

This is an alternative basic recipe for coupling from the past, with the attraction that it can apply where the updating densities are known only up to a multiplicative constant, as is usual in Bayesian MCMC.

We suppose that the conditional update kernel is $f(y \mid x) = k(x)g(y \mid x)$ where the unnormalised density $g(y \mid x)$ is known, but $k(x)$ is unknown, and that there exists an envelope function $h(y)$ such that $g(y \mid x) \leq h(y)$ for all $x, y \in \chi$. We assume $\nu = \int h(y)dy < \infty$, which is quite a demanding condition in an unbounded state space. Finally, we assume that we have a way of sampling from the density $h(y)/\nu$.

The idea is a simple adaptation of rejection sampling: we sample repeatedly under the graph of $h(y)$, yielding i.i.d. points $(Y^{(j)}, V^{(j)}h(Y^{(j)}))$ with $V^{(j)} \sim U(0,1)$ independent of $Y^{(j)} \sim h(\cdot)/\nu$. Let $A_j = \{x \in B_{t-1} : g(Y^{(j)} \mid x) \geq V^{(j)}h(Y^{(j)})\}$: these are the values of x such that $X_t = Y^{(j)}$ is a valid update from $X_{t-1} = x$. Continue to the smallest J such that $\cup_{j=1}^{J} A_j \supset B_{t-1}$. Then use $\Phi(B_{t-1}, U_t) = \{Y^{(j)}, j = 1, 2, \ldots J\}$. (Note that each U_t is a sequence of pairs $(Y^{(j)}, V^{(j)}), j = 1, 2, \ldots.)$

In practice, rather than accumulate the union of the A_j, we check the coverage of χ using a lower squeezing function. And again, partitioning of χ is possible, and is useful in speeding up coalescence.

3. TOWARDS AUTOMATIC METHODS

The algorithms of the previous section demonstrate the possibility of exact sampling in continuous state spaces, and the examples in Murdoch and Green (1998) show that it works in practice, on toy examples, and for a Gibbs sampler applied to the 'pump reliability' data. However, they are expensive to use in terms of algebraic and programming effort. In this section we turn to methods that seem to us to offer a real prospect of useful algorithms for general purposes.

3.1. *Random walk Metropolis*

Random walk Metropolis is a 'vanilla' MCMC method for Bayesian computation, being very easy to code, and usually offering reasonable performance at the expense of a few pilot runs to choose proposal spread parameters.

It is promising for adaptation to CFTP since we can separately couple the proposals and the acceptance decisions; the difficult continuous part can be done once and for all, in a model-independent way, for a proposal distribution chosen at our convenience.

Thus the general objective for a given symmetric proposal density q is to first find a random walk update function ψ such that $\psi(x, U) \sim q(\cdot - x)$ and such that $\psi(B, U)$ is finite with high probability, and then use

$$\phi(B, U) = \bigcup_{x \in A(U)} \psi(x, U^{(2)}) \cup (B \setminus A(U))$$

$$A(U) = \{x \in B : U^{(1)} < \pi(\psi(x, U^{(2)}))/\pi(x)\}$$

is the set of states for which the proposed update is accepted. In practice we work with a $\Phi(B, U)$ for which both terms of the union may be 'rounded up' for convenience.

In the next subsection, we discuss a method for coupling the proposals, and in Subsection 3.3 show how the acceptance/rejection decisions might be automated.

3.2. *The Bisection Coupler*

The bisection coupler is a general purpose random walk coupler for a state space that is a rectangle in any number of dimensions. Suppose first that $\chi = [0, 1]$, and that $q(\cdot)$ is unimodal and symmetric. Let

$$r_{xy}(\cdot) = \min\{q(\cdot - x), q(\cdot - y)\}$$

denote the intersection between two shifted copies of q centered at x and y. Then

1. $r_{xy}(\cdot) = \min_{z \in [x,y]} q(\cdot - z)$
2. $r_{xy}(\cdot) = r_{0,y-x}(\cdot - x)$;
3. $r_{0y}(\cdot) \nearrow q(\cdot)$ as $y \searrow 0$;
4. $r_{xy}(\cdot)$ is symmetric about $(x + y)/2$.

These properties allow a very nice specialisation of the partitioned multigamma coupler. It turns out that we can dissect the area under $q(\cdot)$ into pieces according to the overlaps with $q(\cdot - 2^{-N})$, $N = 0, \ldots$, and then represent $q(\cdot - x)$, for any $x \in [0, 1]$, by rearranging these pieces. The piece created by the N^{th} intersection, $N = 0, 1, 2, \ldots$, occurs in exactly 2^N places over all these decompositions. This is illustrated in Figure 1.

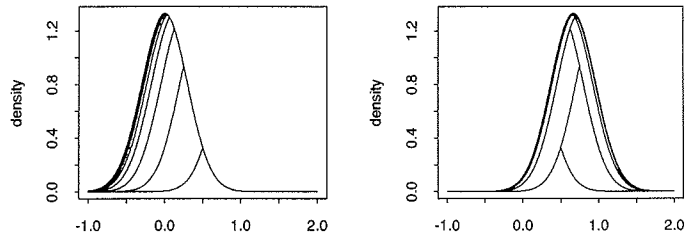

Figure 1. *Illustrating the bisection coupler: the decompositions of the $N(0, 0.3^2)$ and $N(2/3, 0.3^2)$ densities exploited by the algorithm.*

The update $\psi(B, U)$ then consists of sampling one point under the graph of q using U, finding the abscissa in the corresponding piece under $q(\cdot - x)$ for all $x \in B$ by reflections and translations, and setting $\psi(B, U)$ to be the resulting set of abscissae. There will be at most 2^N values in this set (maybe fewer as we do not need to store values outside χ, or moves from outside B), where N indexes the piece into which the random point fell.

We can write all of this as $\psi(x, U) = 2k/2^N - Y$ for $x \in [(2k-1)/2^N, 2k/2^N)$ and $\psi(x, U) = 2k/2^N + Y$ for $x \in [2k/2^N, (2k+1)/2^N)$. Here $U = (V, Y)$, where $V \sim U(0, 1)$ and $Y \sim q(\cdot)$, independently, N is the smallest integer such that $Vq(Y) < q(Y - 2^{-N})$, and k is an integer in the range $0, \ldots, 2^{N-1}$. While N is almost surely finite, and has finite expectation, we find that, when $q(\cdot)$ is continuous at the origin, for large n we have $P(N > n) \approx 2^{-n} q(0)$ so that $E(2^N) = \infty$: the set $\psi(B, U)$ is always finite but can be large. We adopt our usual trick of storing only an interval superset if the list is too long.

If we allow N and k to be any integer values, the bisection coupler may be used on the whole real line. Provided B is bounded, $\psi(B, U)$ will be finite. To allow positive and negative values of x to coalesce, we give a random uniform shift to the coordinate system before partitioning takes place.

Bisection coupling easily generalizes to finite dimensional χ, provided $q(\cdot)$ is unimodal and symmetric about 0 in each of its coordinates: we replace the sequence 2^{-N}, $N = \ldots, 0, 1, \ldots$ by a sequence of vector displacements in which each coordinate is bisected in turn.

3.3. *Automatic Cell Bounds*

For a real impact on the practice of Bayesian computation, it will be necessary for exact simulation to be much more routinely implemented. What are the prospects for automated coupling of the accept/reject decisions?

A minimal general class of models of interest would be the Bayesian hierarchical models based on 'standard' distributions, and built on directed acyclic graphs, such as those implemented in BUGS (Gilks, *et al.* , 1994). Such graphs represent conditional independence statements: given the values $x_{pa(v)}$ of its parents, each variable x_v is conditionally independent of all non-descendants, giving the factorisation:

$$\pi(x) = \prod_{v \in V} \pi(x_v \mid x_{pa(v)}).$$

In turn, this implies a 'near-separability' property that we can exploit. Taking logarithms, we can bound the log-density in arbitrary cells of the parameter space:

$$\sup_{x\in A} \log \pi(x) \leq \sum_{v\in V} \sup_{x\in A} \log \pi(x_v \mid x_{pa(v)})$$

$$\inf_{x\in A} \log \pi(x) \geq \sum_{v\in V} \inf_{x\in A} \log \pi(x_v \mid x_{pa(v)}),$$

and these bounds should not be too slack: in typical models, each variable x_v appears in only a few terms.

For rectangular cells A, and 'standard' distributions, both $\sup_{x\in A} \log \pi(x_v \mid x_{pa(v)})$ and $\inf_{x\in A} \log \pi(x_v \mid x_{pa(v)})$ are explicitly available, typically requiring some tedious one-off calculus on 2 or 3 variables. The bounds apply equally to posterior distributions: observed variables are fixed, and we can ignore the fact that the whole joint density should be re-normalised.

The bounds provide the raw ingredients for partitioned random walk Metropolis coupling, permitting economical checking of whether a proposed update is to be accepted. For example, suppose $A \cap B_{t-1}$ is proposed to be updated to $\{Y_j : j = 1, 2, \ldots, J\} = \psi(A \cap B_{t-1}, U_t^{(2)})$. Then if

$$U_t^{(1)} < \frac{\pi(Y_j)}{\inf_{x\in A\cap B_{t-1}} \pi(x)}$$

we include Y_j in B_t, and if

$$U_t^{(1)} > \frac{\pi(Y_j)}{\sup_{x\in A\cap B_{t-1}} \pi(x)}$$

for any j, we include all of $A \cap B_{t-1}$ in B_t. The set B_t assembled in this way certainly includes $\phi(x, U_t)$ for all $x \in B_{t-1}$.

3.4. *Example: Dirichlet Priors and Multinomial Data*

To avoid problems with unbounded state spaces, we will illustrate the working of the bisection coupler on various target distributions with state space $[0, 1]^d$. We use exact bounds on the target density in each cell.

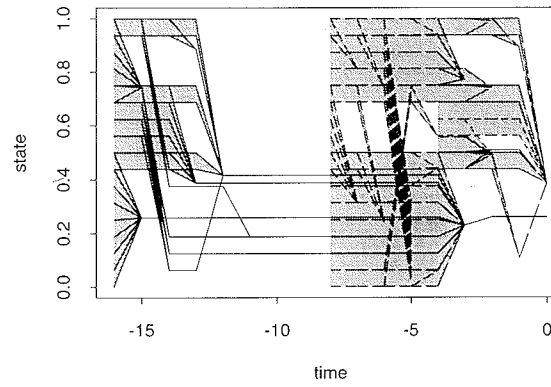

Figure 2. *A single run of the random walk Metropolis bisection coupler with a Beta*(25, 75) *target.*

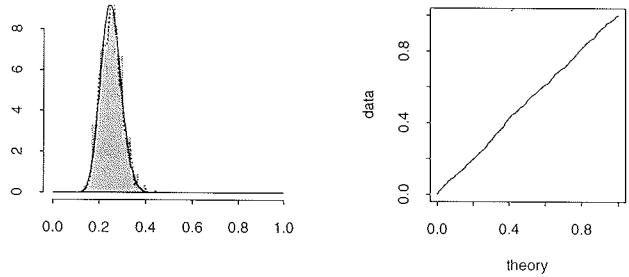

Figure 3. *Histogram and PP-plot of 1000 replicates of the random walk Metropolis bisection coupler with a Beta*$(25, 75)$ *target.*

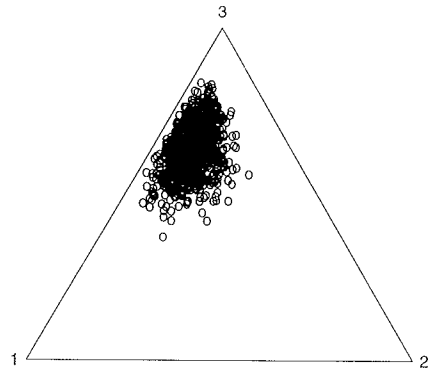

Figure 4. *Barycentric scatter plot of 1000 replicates of the random walk Metropolis bisection coupler for the posterior in the small ABO dataset.*

We begin with a one-dimensional example, with target Beta$(25, 75)$. Figure 2 shows a single run of the random walk Metropolis bisection coupler, using 16 equal-sized cells, and a $\mathrm{N}(0, 0.3^2)$ proposal distribution. Coalescence to a single state (the value 0.2589) was achieved with $M = 16$. The figure should be self-explanatory: grey rectangles show where a proposal was rejected for at least one state in a cell, triangles where such a proposal was accepted, and lines denote simple state-to-state updates after the 'live' states in a cell have been reduced to a finite list; solid lines are used for the successful pass with $M = 16$. In Figure 3, we show a histogram and PP-plot for the final states from 1000 independent replicates of this coupler. The P-value for a Kolmogorov-Smirnov goodness of fit test to the specified target is 0.359. The frequency distribution for M over the 1000 replicates was $(49, 74, 166, 280, 273, 140, 18)$ for the values $(4, 8, 16, 32, 64, 128, 256)$.

Now we turn to a small Bayesian example. In the ABO blood group system, the alleles A and B are co-dominant to O, so under Hardy-Weinberg equilibrium, the phenotypes (A,B,AB,O) occur with probabilities $(p^2 + 2pr, q^2 + 2qr, 2pq, r^2)$ when the gene frequencies for (A,B,O) are (p, q, r) $(p + q + r = 1)$. Given observed phenotype frequencies (9,3,1,10) respectively, we compute the posterior for (p, q, r) under a Dirichlet$(1, 1, 1)$ prior. Note that although we assume multinomial sampling, this is not a conjugate set-up. Figure 4 shows a barycentric plot for the final states from 1000 replicates of a random walk Metropolis bisection coupler for this target. We used 64 equal-sized cells in the (p, q) plane, and the proposal distribution was bivariate circular normal, with standard deviation 0.2.

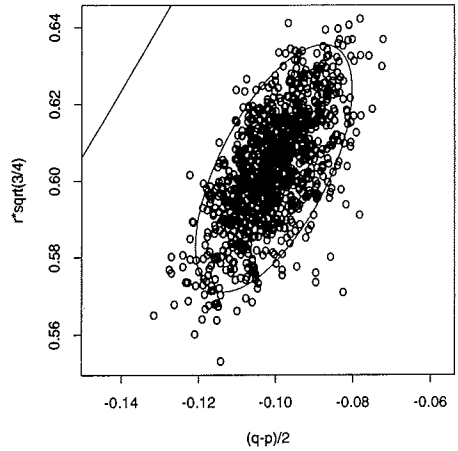

Figure 5. *Magnified view of part of barycentric scatter plot of 1000 replicates of the random walk Metropolis bisection coupler for the posterior in the large ABO dataset, with an elliptical 90% credible region.*

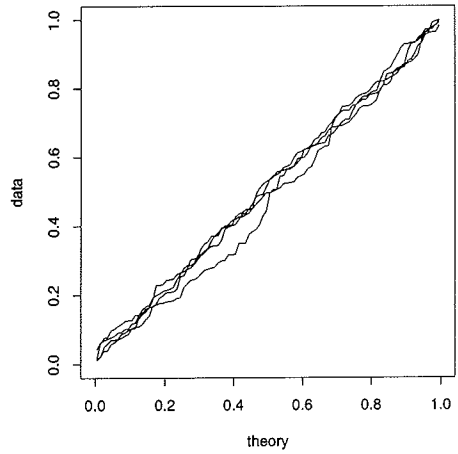

Figure 6. *Superimposed PP-plots of each margin of the output of a random walk Metropolis bisection coupler for a Dirichlet*$(4, 4, 4, 4)$ *target, based on 100 replicates.*

Figure 5 shows results from a similar analysis, but based on the larger data set (179,35,6,202).

Finally, in Figure 6, we show PP-plots of the results of 100 replicates of the coupler for a Dirichlet$(4, 4, 4, 4)$ target: this was implemented using 64 cells in the 3-dimensional parameter space, with $N(0, 0.5^2)$ proposals.

4. DOMINATED CFTP

4.1. *Random bounds on couplers*

Kendall (1998) described a trick which allows CFTP to be used on Markov chains which are not uniformly ergodic. The idea is based on the fact that a stationary realization of the chain will tend to stay in regions of high density. We quantify this by constructing a coupled *dominating* chain which has a stationary distribution with heavier tails than the target, and bound the possible values of the target by the (more easily simulated) values of the dominating chain, thereby reducing the state space to a manageable size.

To make this concrete, consider a one dimensional real-valued chain X_t taking positive values. We construct a dominating chain C_t with the following properties:

Coupling: Both chains are updated together, using the same random inputs. The updating rule can be written as $(X_{t+1}, C_{t+1}) = [\phi_1(X_t, U_{t+1}), \phi_2(C_t, U_{t+1})]$.

Dominance: If $x \leq c$ then $\phi_1(x, u) \leq \phi_2(c, u)$ for all u.

Convenience: C_t has a known limiting distribution from which we can sample directly.

Practical reversibility: We know how to run the bounding chain in reverse, i.e. when the dominating chain is in equilibrium, we know the conditional distribution of (C_{t-1}, U_t) given C_t and can sample from it.

In addition to these conditions, we need to ensure that the chain X_t eventually falls below C_t regardless of its starting value. For the cases of interest to us, coupled X_t chains will eventually coalesce, so this is automatic.

We then proceed as follows. To obtain the set B_{-M} of possible values of X_t at time $t = -M$, we sample C_0 from its stationary distribution, and then generate pairs (C_{t-1}, U_t), $t = 0, \ldots, -M + 1$. Since we know that any X_t path, if run from far enough back in time, would be caught below C_t by time $-M$, we set B_{-M} equal to $[0, C_{-M}]$. Using ϕ_1 and the just-generated U_t values, we carry it forward to time 0 for the usual CFTP test of coalescence.

The beauty of this dominated CFTP algorithm is that it does not require uniform ergodicity of the Markov chain on the whole state space. We only need to coalesce values on the random sets below C_t. Most chains that would arise in practice are uniformly ergodic on bounded sets, so coalescence may be achievable in practice.

4.2. *Applying Dominated CFTP to Random Walk Metropolis*

Random walk Metropolis chains on unbounded state spaces are not uniformly ergodic, so standard CFTP will not work for them. Here we describe some details of implementing dominated CFTP for these chains.

Suppose that our target distribution has density proportional to $\pi(\cdot)$, and that π has positive support. Furthermore, π has a right tail that is stochastically dominated by a known distribution; for illustration we'll assume an exponential tail. Specifically, there exist constants $C > 0$ and $D > 0$ such that for $0 < d < D$, $C + D \leq x_1$, and $C \leq x \leq x_1$,

$$\frac{\pi(x + d)}{\pi(x)} \leq \frac{\pi_c(x_1 + d)}{\pi_c(x_1)}$$

where $\pi_c(x) = \exp(-x)$ for $x > C + D$ is proportional to the density of an Exponential(1) random variable shifted $C + D$ units to the right.

What this condition is designed to achieve is the following. We replace U_t in the previous section with the pair (D_t, U_t), where $D_t \sim U(-D, D)$ and $U_t \sim U(0, 1)$ are independent. Then random walk Metropolis for the target process may be written as $X_{t+1} = \phi_1(X_t, D_{t+1}, U_{t+1})$

where $\phi_1(x, d, u) = x + d$ if $\pi(x + d)/\pi(x) > u$ and x otherwise. Similarly, we implement the coupled dominating process as $C_{t+1} = \phi_2(C_t, D_{t+1}, U_{t+1})$, where we accept a move if $\pi_c(C_t + D_{t+1})/\pi_c(C_t) > U_{t+1}$.

With these definitions, the required ordering of X_t and C_t holds. Provided $X_t < C_t$, if X_t accepts a move upwards, so will C_t (unless X_t is so far below C_t that the paths could not cross), and similarly moves of C_t downwards imply corresponding moves of (close enough) X_t values. The other requirements of dominated CFTP are also met. In particular, to simulate the dominating chain backwards, we sample $\tilde{U} \sim U(0, 1)$ and $\tilde{D} \sim U(-D, D)$; then if $\tilde{U} > \exp(\tilde{D})$ or $C_{t+1} - \tilde{D} < C + D$, we set $C_t = C_{t+1}$, $D_t = -\tilde{D}$ and $U_t = \tilde{U}$, and otherwise set $C_t = C_{t+1} - \tilde{D}$, $D_t = \tilde{D}$ and $U_t = \tilde{U}/\exp(\tilde{D})$.

One problem with this formulation is that it couples chains by proposing the same D_t value for every state. But this coupler does not coalesce, so the dominated CFTP algorithm would never terminate. We need to incorporate a coalescing update rule, such as the bisection coupler, but the bisection coupler generally will not respect the ordering: it uses different steps D_t from different states, and often leads to crossing paths.

A solution to this problem is to use a method like that of Corcoran and Tweedie (1998): construct an update formula which coalesces for some states but not all. They used a multigamma coupler for central states and achieved global monotonicity, and we could do the same. However, global monotonicity is not needed: provided $X_t \leq C$ and $D_t \in [-D, D]$, we do not care whether $\phi_1(X_t, D_{t+1}, U_{t+1})$ is monotonic. We only need to preserve the condition $X_{t+1} \leq C_{t+1}$, and it is automatically preserved by the bound on D_t. Thus we can use the bisection coupler on the states below C, and use common D_t proposals on other states.

4.3. *Extensions to More General Targets*

The extension of dominated CFTP to two-tailed distributions is straightforward. We can put bounding processes on both the upper and lower tails, i.e. define coupled processes C_t^0, X_t and C_t^1 by $(C_{t+1}^0, X_{t+1}, C_{t+1}^1) = (\phi_0(C_t^0, U_{t+1}), \phi_1(X_t, U_{t+1}), \phi_2(C_t^1, U_{t+1}))$ such that if $c_0 \leq x \leq c_1$ then $\phi_0(c_0, u) \leq \phi_1(x, u) \leq \phi_1(c_1, u)$ for all u. The only new twist is that we need to sample from the joint limiting distribution of the pair of bounds at time 0, and update both backwards.

However, if we can use the same bounds on each tail then this is very easy. Provided we choose our random walk coupler to make jump proposals $-D_t$ for states less than $-C$ when it makes proposals D_t for states greater than C, we will have an equilibrium distribution with $C_t^0 = -C_t^1$ and only C_t^1 need be simulated.

The extension to p-dimensional processes is not so easy. If we could put separate bounds on each coordinate and these bounds applied uniformly across the values of the other coordinates, then we could bound each coordinate independently and we would not need anything new to sample at time 0 or to evolve the bounds backwards in time. However, most interesting p-dimensional distributions do not allow for uniform bounds. For instance, among the multivariate normals, uniform bounds imply zero correlation between the coordinates: not a very interesting case! It may turn out that we can transform the space to allow for uniform bounds as we can in the multivariate normal case, but it remains to be seen whether these transformations will be widely available.

4.4. *Example: Dominated CFTP for a Univariate Case*

To illustrate the dominated CFTP algorithm, we use it to simulate a value from a standard Gaussian distribution. Ratios of the tails of the target density are bounded by ratios of e^{-x} for $x > C = 1$; we chose to use $D_t \sim U(-2, 2)$ random step proposals. This gave an

Exponential(1), shifted 3 units to the right, as the upper bound, and its negative as the lower bound. We used the modified bisection coupler on the range $[-1, 1]$ to induce coalescence in the target paths.

We made one other modification to the algorithm. Instead of testing whether the whole interval $[-C_{-M}, C_{-M}]$ coalesces, we do a first pass test of just the two points $-C_{-M}$ and C_{-M}. If those two do not coalesce, it is unnecessary to test any others. Simulating just two paths makes the calculations much faster than simulating all paths.

A single run of this coupler is illustrated in Figure 7. It may be seen that in this run, the bounding points first coalesced when $M = 8$. However, when the full interval was run, there were still 4 possible values of the target process at time 0, so $M = 16$ was attempted. All values from that run coalesced at time $t = -12$; this path was carried forward to time 0, where it yielded the simulated $N(0, 1)$ value -0.755.

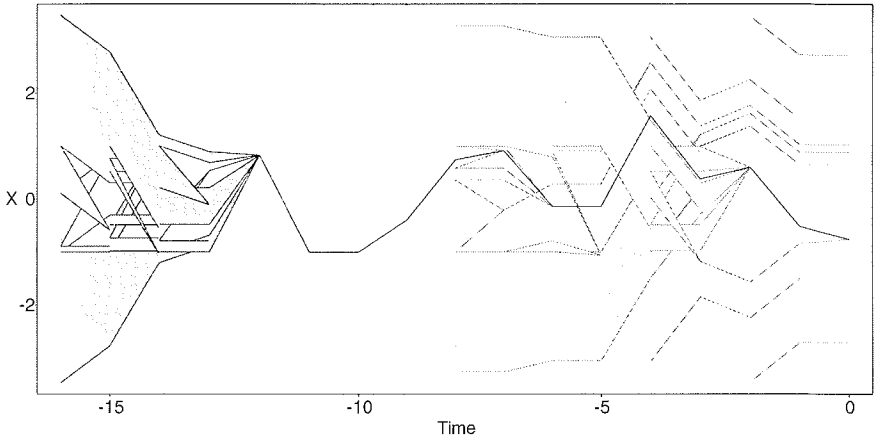

Figure 7. *A single run of the dominated CFTP algorithm with a $N(0, 1)$ target.*

5. CONCLUSIONS

At the start of this paper we expressed optimism about the eventual routine adoption of exact sampling methods in Bayesian computation, and we hope we have communicated some of this to the reader. In this section we will outline some of the remaining work to be done.

To date, exact sampling has been most spectacularly successful in its applications to models of very particular structure in many dimensions, where convergence diagnostics for MCMC methods are unavailable or unreliable. On the other hand, the methods for rather general models described in this paper work best in few dimensions: even the three dimensional Dirichlet example described in Section 3.4 was a stretch, taking much longer to run than a standard simulation from a Dirichlet distribution. To merit routine use, methods are needed that work in the middle case, from a few to a few dozen dimensions.

Efficiency of use of the calculations is an important issue. A great deal of effort goes into producing just one sample from the target distribution. In some of the statistical physics and spatial process applications of exact sampling, one sample is already highly informative due to spatial ergodicity, but that perspective is of no use in Bayesian statistics. We have to decide how to use the opportunity to draw an exact sample in the MCMC context: do we make one (expensive) exact draw as a starting point for one long MCMC run, accepting that

our ultimate sample will be dependent but is guaranteed to have the correct distribution? Or do we repeatedly draw exact samples, avoid ordinary MCMC and escape dependence as well as convergence detection? Presumably, the optimal choice will lie somewhere in the middle, and will depend on the mixing time of the chain, the theoretical coupling time of the structure function ϕ, any inefficiency in computer time in replacing ϕ by a convenient Φ, and the cost of evaluating the functionals of interest from the generated states.

If the time spent on exact draws is an appreciable fraction of the total simulation time, then surely better use of the information gleaned during the search for coalescence could be made than to throw it all away; it must be of some value, at least in tuning the CFTP implementation, and possibly more directly. If such information can be exploited, then it will certainly make the effort of exact sampling more attractive.

There is a certain aesthetic appeal to truly exact sampling, but when these methods are not feasible, simulations could still be improved by using ideas from CFTP. For example, Johnson's (1996) idea of coupling a finite set of paths forward in time to identify burn-in times might be improved if it were used to specify the initial set B_{-M} in a CFTP algorithm: then the bias induced by stopping based on a coalescence time would be removed.

ACKNOWLEDGEMENTS

We are grateful to Laird Breyer and Jeffrey Rosenthal for helpful discussions about dominated CFTP. This research was supported in part by EPSRC and NSERC Research Grants to the respective authors.

REFERENCES

Asmussen, S., Glynn, P. W. and Thorisson, H. (1992). Stationary detection in the initial transient problem. *ACM Transactions on Modeling and Computer Simulation* **2**,130–157.

Corcoran, J. N. and Tweedie, R. L. (1998). Perfect sampling of Harris recurrent Markov chains. *Tech. Rep.*, Colorado State University.

Fill, J. A. (1998) An interruptible algorithm for perfect sampling via Markov chains. *Ann. Appl. Prob.* **8**, 131–162.

Fismen, M. (1997). Exact simulation using Markov chains. *Tech. Rep.* **6/98**, Department of Mathematics, Norwegian University. (available at `http://www.math.ntnu.no/preprint/statistics/1998`)

Foss, S. G. and Tweedie, R. L. (1998). Perfect simulation and backward coupling. *Stochastic Models* **14**, 187–203

Gilks, W. R., Thomas, A. and Spiegelhalter, D. J. (1994). A language and program for complex Bayesian modelling. *The Statistician* **43**, 169–178.

Häggström, O., van Lieshout, M. N. M., and Møller, J. (1999). Characterisation results and Markov chain Monte Carlo algorithms including exact simulation for some spatial point processes. *Bernoulli* (to appear).

Häggström, O. and Nelander, K. (1997a). Exact sampling from anti-monotone systems. *Statistica Neerlandica*. (to appear).

Häggström, O. and Nelander, K. (1997b). On exact simulation of Markov random fields using coupling from the past. *Scandinavian J. Statist.* (to appear).

Johnson, V. E. (1996). Studying convergence of Markov chain Monte Carlo algorithms using coupled sample paths. *J. Amer. Statist. Assoc.* **91**, 154–166.

Kendall, W. (1998). Perfect simulation for the area-interaction point process. *Probability Towards 2000*. (L. Accardi and C. C. Heyde, eds.) New York: Springer, 218–234.

Lindvall, T. (1992). *Lectures on the Coupling Method*. Chichester: Wiley.

Møller, J. (1999). Perfect simulation of conditionally specified models. *J. Roy. Statist. Soc. B* **61** (to appear).

Murdoch, D. J. and Green, P. J. (1998) Exact sampling from a continuous state space. *Scandinavian J. Statist.* **25**, 483–502.

Propp, J. G. and Wilson, D. B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms* **9**, 223–252.

Propp, J. G. and Wilson, D. B. (1998). How to get a perfectly random sample from a generic Markov chain and generate a random spanning tree of a directed graph. *Journal of Algorithms* **27**, 170–217.

DISCUSSION

PAUL DAMIEN *(University of Michigan Business School, USA)*

Based on the scheme of *coupling from the past* (CFTP) the notion that convergence should no longer be an issue in routine MCMC applications appears remarkably seductive. This may sound overly radical, but, at the very least, the strategic alliance of coupling and regeneration schemes can help identify burn-in times while using MCMC methods; i.e., it can serve as a powerful convergence diagnostic tool: this is a substantial advance; Valen Johnson discusses and exemplifies this alliance in his recent JASA paper.

If *exact sampling* methods (interchangeable with the phrase *backward sampling*) are to be successful in practice, then they should work in continuous state spaces. And this is what the authors have done in this paper, as well as a companion paper that is to appear in the Scandinavian Journal. In the continuous setting, the trick is to *induce* discreteness, which the authors address in detail.

In the Gibbs sampling context, the CFTP scheme can be used to construct rejection or Metropolis-Hastings couplers to simulate conditional densities; a nice feature of these couplers is that the density has to be known only up to proportionality, unlike the multigamma coupler.

In the Scandinanvian Journal paper, the authors reanalyse the well-known (Gelfand & Smith, 1990) pump model/data by constructing and contrasting different couplers. For this 11 parameters model, a two-component Gibbs sampler comprises full conditional distributions that are both gamma distributions, where one component, of course, is a 10 dimensional vector. Two conclusions are striking from the authors' analysis. First, the percentage of times that different sample paths coalesce and the acceptance rates of the different couplers critically depend on the magnitude of the inverse scale parameter; this type of a constraint is somewhat unattractive if exact sampling methods are to be labelled "general purpose algorithms." Secondly, the pump model is well-behaved. I suspect that the type of couplers entertained by the authors would lead to much slower acceptance rates (leading to lower coupling percentages) in more complicated models. Thus the question "has the MCMC chain converged?" is replaced by "how far back should one go to ensure that the chains coalesce by time zero?"

In (conventional) *forward sampling*, we confront many problems in implementing rejection methods and/or Metropolis-Hastings type algorithms: as examples, the need to find dominating densities and/or perform awkard supremum calculations. It appears that these problems may be difficult to obviate within the context of backward sampling as well.

All diagnostic protocols fail to satisfactorily assess convergence in the analysis of finite mixture models with varying number of components. Consider an m-component mixture model. Typically, convergence for the parameters in the model are tested after m has reached stationarity for fixed m-values. However, it is well-known that even after a long run some values of m will not be visited often; convergence diagnostics are of little comfort in these cases.

Since this has been a principal area of research of one of the authors, Professor Green, do the authors have any insights as to whether the CFTP idea can be extended or modified to provide a satisfactory answer to the finite mixture modelling problem?

The authors are to be: thanked for unveiling and describing new ideas from inter-disciplinary literature; congratulated for developing and illustrating these ideas in a lucid manner to the statistical world. Understandably, they have adopted a cautiously optimistic tone regarding the future of exact sampling on statistical practice. But the fundamental concept underlying exact

sampling is as compelling as the Gibbs sampler. Learning from that experience, then, it is tempting to jettison the authors' caution into the Mediterranean.

LUKE TIERNEY *(University of Minnesota, USA)*

Coupling from the past is a fascinating and valuable idea, and I congratulate the authors on their important work to apply this idea in continuous state spaces.

The basic CFTP idea is still quite new and may be worth extending and modifying in a number of directions. The structure of the partitioned multi-gamma coupler suggests one possible direction that might allow more flexibility in the generation of the paths. In the partitioned multi-gamma coupler the chains contain an embedded renewal sequence corresponding to the times when the R_i distributions are used. These times can be simulated directly as cumulative sums of geometric-ρ variables. The tours of a chain between these times are conditionally independent given the times and the partition cells containing the chain at these times. The tours can be generated by generating initial states from the R_i for each partition set and then generating new states from the appropriate Q_i transition kernels until the next renewal time. Chains that find themselves in the same partition set at the next renewal time are coalesced. It is possible, but no longer necessary, to use identical random variables to generate tours for all chains. Using identical renewal times and proper marginal behavior of the tours is sufficient for the coupling argument to apply. Using identical variates may allow additional coalescing to occur between renewals. On the other hand, generating the tours independently may allow more flexibility in generation strategies. Independent generation may also make it easier to compute tours started from different partition sets in parallel since the need for synchronization on common variates is eliminated. Similar arguments may apply to other couplers.

Another observation that applies to the partitioned multi-gamma sampler is that the tours from the first backward renewal time to zero provide candidates for the value of a draw from the target distribution; the remaining tours serve to eliminate candidates until only one is left. In some settings it may be sufficient to produce these candidates as starting values for chains run forward from time zero. A mixture of this finite set of chains is known to have equilibrium distribution π; considering all possible mixtures may produce bounds that are useful in practice. Again this observation may apply to other samplers as well.

A drawback of coupling methods available to date is that they all seem to require rather detailed analysis of the target distribution. To keep things in proper perspective, it is important to examine what other strategies are available at a comparable price in analytic effort. It would appear that in all the examples given by Green and Murdoch similar calculations to those needed to set up a coupler can also be used to set up a rejection sampler for the target distribution. Coupling is in principle a more general strategy, but to make the case that it also more general in practice it would be useful to have some examples of Bayesian inference settings where coupling ideas can clearly produce a method for drawing from the target distribution more easily than rejection sampling.

A possible drawback of exact sampling methods based on CFTP is that, like importance sampling, rejection sampling, and regenerative analysis methods, they are global in the sense that they need to be applied to an entire target distribution—they are not really useful when applied to conditional distributions in a larger MCMC framework. Sampling methods, in particular importance sampling, were used quite extensively for Bayesian inference in the 1980's with some success, but sampling methods did not become widely used until the introduction of MCMC methods. There are many reasons for this, but one of the most important factors, in my view, is that the idea of conditioning is a divide and conquer strategy: Larger problems can be decomposed into simpler ones by conditioning, and solutions for sampling the simpler conditional problems can be combined into solutions for the original problem using the Markov

chain framework. New sampling methods need to be examined in this light. A method that can be applied usefully to conditional distributions is more likely to be useful in a wide range of problems than one that only applies to the entire target distribution. It would of course be possible to sample conditional distributions by CFTP, but the analysis effort would need to be reduced significantly. Even then, the advantage of exact sampling as part of an MCMC over a single or fixed number of steps from a Markov chain sampler of the conditional distribution is questionable.

Another point to keep in mind, as Green and Murdoch point out, is that the value of being able to draw from the exact target distribution has to be considered in the context of the cost of doing so. In many problems strategies for exact sampling, such as rejection sampling, have been available for a long time but are usually considered not worth the effort. The cost can be reduced by using a possibly expensive exact draw from the target as a starting point for a cheaper Markov chain. But a peculiarity of this approach as an approach to practical Bayesian inference is that it only makes sense from a frequentist perspective: if you condition on your starting value, you are back with a non-stationary chain.

Despite these minor concerns, this work represents a very significant contribution and I again congratulate the authors on their excellent efforts.

PHILIP J. EVERSON *(Swarthmore College, USA)*

It is worth noting that the *stochastic recursive sequence* representation $X_{t+1} = \phi(X_t, U_{t+1})$, referred to by Green and Murdoch in (3.4), is more general than it may appear at first glance. Take for example the simple problem of generating bivariate draws from a Uniform triangle: $f(x, y) = 2$, for $0 \leq x \leq y \leq 1$. A Gibbs sampler for this problem updates $\{X, Y\}$ pairs by sampling $X_{t+1} \sim \text{Unif}\,(0, y_t)$, $Y_{t+1} \sim \text{Unif}\,(x_{t+1}, 1)$. This appears to violate the requirement that U_t be a sequence from some *fixed* distribution, as the bounds on X and Y change at each step. However, we can always think of U_t as representing the sequence of seeds driving the underlying random number generator. If, after M steps, random sequences started with the same seed from all initial conditions have coalesced, then certainly the initial conditions are irrelevant from that point on.

An alternative to running each sequence for M steps before checking coalescence, is to instead update every sequence one step at a time. Consider a discrete analog to the above problem, where X and Y take on values $0.0, 0.1, \ldots, 1.0$. The following Splus code will update any one sequence, using the "same" random numbers:

```
old.seed <- .Random.seed
if (y==0) {x<- 0}; else
x <- sample(seq(0,y,.1),1)
if (x==1) {y<- 1}; else
y <- sample(seq(x,1,.1),1)
new.seed <- .Random.seed
.Random.seed <- old.seed
```

After advancing every sequence by one step, the next update is begun using new.seed, and the algorithm continues until all sequences have coalesced. With a starting seed of 1, sequences begun at any of the 66 allowable $\{X, Y\}$ pairs will take on the value $\{0.1, 0.7\}$ after only two steps. They of course remain coupled from that point on, indicating that the sampler has achieved equilibrium.

CHRISTIAN P. ROBERT *(CREST, Insee, Paris, and Université de Rouen, France)* and D.M. TITTERINGTON *(University of Glasgow, UK)*

This paper greatly widens the scope of perfect simulation in continuous settings in that it does not require renewal techniques as in Murdoch and Green (1998). The authors must therefore be congratulated for opening new and broader entrances to this field. As in every partitioning method, there is still a difficulty with higher dimensions in that the number of terms involved in the bisection coupler is likely to grow quite rapidly with the dimension.

Following Propp and Wilson (1996), the authors adopt a geometric increase in the backward steps. Is there any theoretical ground for this choice, as opposed to the naive linear increase (1,2,3,...)?

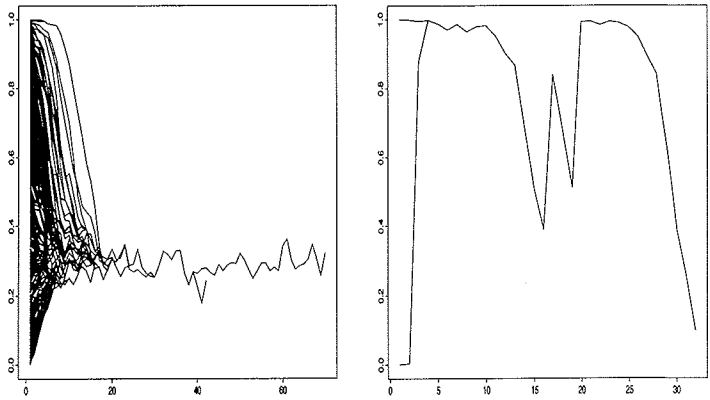

Figure 1. *Comparison of an exhaustive (left) and of a monotone (right) CFTP on the chain* $(p^{(t)})$ *in the case of a two component mixture of normals with 495 observations. (The number of backward steps is* 69 *in the exhaustive case and* 32 *in the monotone case.)*

Although the authors do not consider the possibility of monotone CFTP in continuous setups, there are settings where monotonicity can speed up implementation and reduce computing time. For instance, consider a two-component mixture

$$pf_1(x) + (1-p)f_2(x)\ , \tag{1}$$

where both f_1 and f_2 are known and $p \sim \mathcal{U}(0,1)$. As shown by Hobert *et al.* (1998), CFTP can be implemented in this setting by considering only $n+1$ coupled parallel chains (instead of a continuum on p), if n is the number of observations. However, when n grows, the number of backward steps also increases. Since a Gibbs move from p to p' involves generating n uniforms u_i, then computing

$$n_1 = \sum_i \mathbb{I}_{u_i \leq pf_1(x_i)/\{pf_1(x_i)+(1-p)f_2(x_i)\}}$$

(which can only take $n+1$ values, thereby determining the number of p chains to be considered), and deriving $p' = \Sigma_{j=1}^{n_1+1} w_j / \Sigma_{j=1}^{n+2} w_j$, where $w_j \sim \mathcal{E}xp(1)$, it is easy to see that starting from $n_1 = 0$ and $n_1 = n$ provides an envelope for all possible chains, since $p_1 \leq p_2$ implies $p_1' \leq p_2'$. Figure 1 provides a comparison of both implementations (with different coupling structures). Extensions to the 3 component case are much more delicate, as revealed in Hobert *et al.* (1998).

REPLY TO THE DISCUSSION

We are grateful to each of the discussants for their thorough reading of both the present paper and our Scandinavian Journal article, for their generous comments, and for their insightful and stimulating discussions. They are evidently as excited as we are by the apparent opportunities provided by coupling from the past, and share our mixed feelings about whether some tremendous benefits for Bayesian computation are waiting to be uncovered, or whether this will all turn out to be just an intriguing sideline.

Four years gestation should be enough for real applications to emerge, so one measure of how it turns out will be the number of papers at the *next* Valencia that exploit exact sampling. Whether this is zero or a lot, we are pleased that we took the somewhat risky decision to present these ideas at the conference.

Most of the discussants allude to the cumbersome algebra and analysis needed to set up coupling from the past in all but the simplest of models. This is a central issue. There is a very real danger that in pursuing the benefits of "guaranteed convergence", implementors will be throwing away the very freedoms that MCMC bought for them in the first place. Any substantive data analysis that is limited by what can be done exactly turns the clock back to when tractability was a dominant issue in Bayesian analysis.

For these reasons, we see partial automation of CFTP implementation as absolutely crucial to the practical application of these ideas. The methods introduced in the Scandinavian Journal article—the multigamma and rejection couplers, and their partitioned versions—are very special. Their existence does little more than demonstrate their existence!

However, the methods introduced in Sections 3 and 4 of the present paper—bisection coupling and dominated coupling—claim to be much more generic. They are constructed from an analysis of the target distribution, which in Bayesian computation is usually explicit if complicated, rather than the transition mechanism of the chosen Markov chain, which is usually extremely awkward. One price to be paid for their generic character is probable increased coalescence time, as in several places in the derivation of these methods, exact expressions are replaced by bounds, and sets of update states enlarged for convenience of representation. Such conservatism does not affect the exactness of the ultimate draw from the target, but may delay coalescence—perhaps, fatally, to time infinity—and complicate analysis of performance.

Perhaps one value of the earlier methods is to shed some light on the roles of regeneration and renewal in coupling. Professor Tierney exploits this in proposing some very nice modifications of the multigamma method. It is potentially useful to be able to relax from the requirement to organise housekeeping of the random number streams to ensure that $X_{t+1} = \phi(X_t, U_{t+1})$. However, we repeat that we do not see multigamma coupling as a general method, although it may play a role as a building block.

We thank Professor Everson for emphasising that the update function/stochastic recursive sequence representation $X_{t+1} = \phi(X_t, U_{t+1})$ allows much flexibility. We exploited somewhat similar formulations in our implementations of rejection couplers, which are set out in the Scandinavian Journal paper. Such flexibility can be taken much further. Taking a finite state space example for clarity, suppose that $\chi = \{1, 2, \ldots, S\}$, and that the $U_t = (U_t^{(1)}, \ldots, U_t^{(S)})$ are S-dimensional. Then it is legitimate to take $X_{t+1} = \phi(X_t, U_{t+1}) = \tilde{\phi}(s, U_{t+1}^{(s)})$ when $X_t = s$, for an appropriate $\tilde{\phi}$, so that we actually use a *different* random number for each current state X_t! We have nevertheless ensured what we must, namely that the update from X_t at time t is the same on every occasion that this update is considered.

For us, the most substantial reason to doubt the eventual widespread use of CFTP for truly exact sampling in Bayesian analysis is the non-modular nature of all existing CFTP prescriptions. Ordinary MCMC methods are all inherently modular, graphical modelling representations of

statistical models are modular, and the practice of model-building for data analysis is modular. We completely agree with Professor Tierney in his analysis of this issue, based on a "divide-and-conquer" interpretation of the usual MCMC recipes. The only place in our methods where we can work in a modular fashion is in constructing the approximate cell bounds described in Section 3.3. We do have some prototype code that can create tables of cell bounds automatically from a BUGS specification of the model, but this is untested in real problems, and the whole strategy of partition by rectangular cells is of course limited by the curse of dimensionality.

The role played by monotonicity in CFTP deserves further study. We have stressed the non-monotone nature of our methods, meaning that we need to use structure functions $\phi(x, u)$ that are not monotonic in x. (Of course, other aspects of monotonicity are being used in establishing the bounds we need; for example, acceptance probability is monotonic in the target density value for the current state, an essential ingredient in Section 3.) Professors Robert and Titterington give a nice example, demonstrating that we were wrong to rule out the possibility of monotonicity with respect to state in continuous-parameter problems, but we are not sure how often such tricks will be available.

In this connection, we should correct an impression given in Section 1.3. Häggström and Nelander (1997b) do consider non-monotonic couplers for Markov random fields. They include an example where use of a monotonic coupler is vastly superior to a non-monotonic one in terms of coalescence time, and further research on such comparisons would be interesting. But, as with Robert and Titterington's example, it is not clear how much of the gain can be attributed to monotonicity (allowing the tracing of fewer paths, 2 instead of $n + 1$ in the mixture example), as the coupling methods are entirely different (one giving intrinsically faster coalescence).

Professor Damien is right to note that the key to creating coalescing paths in a continuous parameter problem is to look for ways of inducing discreteness; this was indeed our motivation. This idea has appeared previously in Nummelin's splitting technique, translated into practical MCMC methodology by Mykland, Tierney and Yu (1995).

We have already begun thinking about exact sampling for variable dimension problems such as mixture analysis, as suggested by Professor Damien. (In fact, we do not necessarily buy his motivation for doing this; we feel that the breakdown of existing diagnostic methods comes not in going from finite to infinite or variable dimensions, but in going from small to moderate dimensionality. In all cases, you will apply diagnostics to some carefully-chosen scalar functionals; all that changes with dimension is the capacity of a few such functionals to capture the behaviour of the entire chain.) It is not too far-fetched to surmise that it might be possible to couple MCMC methods for mixtures. These problems do have some of the character of point-process models—here the points would be marked—and several such models have already been addressed by CFTP. Robert and Titterington's implementation of CFTP for a particular one-parameter mixture problem is a start, although we are not sure how far this example can be extended.

On the question of how to increase the trial values of M in coupling from the past, raised by Professors Robert and Titterington: of course, any increasing sequence is valid, and the matter is simply a question of computing time. Propp and Wilson give an argument to support a geometric increase. Besides that argument, there is the heuristic reasoning that computing a failure is usually much more expensive than computing a success, because many more paths need to be followed, so you want to reduce the number of failures. Since the distribution of T often has a heavy tail, you need a rapidly increasing sequence of trial values.

For the small Beta problem discussed in Section 3.4, we have given the histogram of the successful M values in 1000 independent replications of the random walk Metropolis method. This says something about the distribution of T, since $T \in (M/2, M]$. Clearly a linear increase

in M would tie T down to be a little smaller, but at the cost of many more failed attempts to achieve coalescence. The approximately symmetric shape of the distribution of $\log T$ also seems to support the geometric schedule. Empirically, for this example, the schedules $(16, 32, 64, \ldots)$, $(8, 32, 128, \ldots)$ and $(32, 128, 512, \ldots)$ all take about the same time on average, $(4, 8, 16, \ldots)$ takes 30% longer, but $(8, 16, 24, \ldots)$ takes twice as long, and $(8, 10, 12, \ldots)$ seven times longer.

Some further analysis of the running time implications of various details of CFTP implementation and use will be found in Murdoch and Rosenthal (1998).

These remarks perhaps also answer Professor Damien's question about "how far back should we go to get coalescence by time 0". The point is, of course, that we do not know, initially, but that the CFTP protocol provides a self-tuning mechanism for finding out. Incidentally, if CFTP is independently replicated, it is legitimate to change the M schedule for later runs in the light of experience about when coalescence typically occurs, gained from the early runs.

We now turn to questions about use of CFTP in an approximate mode. It may well be that in moderate- to high-dimensional Bayesian problems, "exact exact sampling" is too hard, and that some less aesthetically pleasing but still useful role for CFTP ideas will be the best we can do, as mentioned at the end of our conclusions. Thus we agree with Professor Damien that combining our methods with other strategies, like Johnson's, is a good idea. Even if exact sampling is impractical the couplers we present in this paper are fairly easy to implement and could be used as diagnostics instead. Non-exact uses of couplers are also discussed in Murdoch and Rosenthal (1998). Regarding Professor Tierney's ideas for a use for non-coalesced endpoints: unfortunately, we do not have a coupler that is both general purpose and easy to apply, where it is straightforward to come up with a finite list of candidates from an easily identifiable time. The rejection and multigamma couplers are too hard to apply, and the random walk couplers can have infinite sets of states that take a long time to move.

Finally, we know this is a Bayesian conference and a Bayesian volume, but we still think it is entirely appropriate to take a frequentist perspective in judging the quality of a Monte Carlo sample!

ADDITIONAL REFERENCES IN THE DISCUSSION

Gelfand, A.E. and Smith, A.F.M. (1990) Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85**, 398–409.

Hobert, J.P., Robert, C.P. and Titterington, M. (1998). On perfect simulation for mixtures of distributions. *Tech. Rep.* **9835**, CREST, INSEE, Paris.

Murdoch, D. J. and Rosenthal, J. S. (1998). Efficient use of exact samples. *Tech. Rep.*, Queen's University, Kingston.

Mykland, P., Tierney, L. and Yu, B. (1995). Regeneration in Markov chain samplers. *J. Amer. Statist. Assoc.* **90**, 233–241.

Propp, J.G. and Wilson, D.B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms* **9**, 223–252.