Moralizing Perfect Sampling

James P. Hobert*
Department of Statistics
University of Florida
Gainesville, FL
jhobert@stat.ufl.edu

Christian P. Robert†
Université Paris Dauphine
and CREST, Insee
Paris, France
robert@ensae.fr

October 2000

*This research partially supported by NSF Grant DMS-00-72827 and by the Centre National de la Recherche Scientifique (CNRS) during a visit to the Centre de Recherche en Economie et Statistique (CREST) in Paris in June 2000.

†This research partially supported by EU TMR network ERB-FMRX-CT96-0095 on *"Computational and statistical methods for the analysis of spatial data"* and by the Department of Statistics at the University of Florida during a visit in August 1999.

AMS 1991 *subject classifications.* Primary 62C15; secondary 60J05

Key words and phrases. Burn-in, Gibbs sampler, Kac's Theorem, L_1-Wasserstein distance, Markov chain, Minorization condition, Mixture distribution, Monte Carlo, Split chain, Total variation distance

Abstract

Let Φ be an irreducible, aperiodic, Harris recurrent Markov chain with invariant probability measure π. We show that if a minorization condition can be established for Φ, then π can be represented as an infinite mixture from which it may be possible to sample. Making exact draws from this mixture (and hence from π) requires the ability to simulate two different random variables, T^* and N_t. When the *small set* in the minorization condition is the entire state space (and hence Φ is uniformly ergodic), simulating these random variables is straightforward and the resulting algorithm turns out to be equivalent to Murdoch and Green's (1998) *multigamma coupler*. Thus, we are able to achieve perfect sampling in the uniformly ergodic case without any appeal to *coupling* or *backward simulation*. When the small set is not the state space, simulating T^* is more problematic. Fortunately, T^* is univariate and has support $\mathbb{N}$. We construct an estimate of the mass function of T^* and an asymptotic bound on the error of this estimate. These results are used to address rigorously the issue of burn-in. Specifically, we form an approximation of π whose error can be controlled and which can be used as an initial distribution for Φ. We illustrate this with a realistic example. Finally, we provide a simple Markov chain Monte Carlo (MCMC) algorithm whose stationary distribution is that of T^*. Finding a perfect sampler based on our MCMC algorithm for the univariate, discrete T^* is tantamount to the ability to sample exactly from π.

1 Introduction

Let Φ be an irreducible, aperiodic, Harris recurrent Markov chain with invariant probability measure π. There are many different ways to express the measure π in terms of the Markov transition kernel of Φ (Meyn and Tweedie, 1993, Chapter 10). We derive a new representation for π that is based on the Markov transition kernel and a minorization condition for Φ. Our representation takes the form of an infinite mixture

$$\pi(A) = \sum_{t=1}^{\infty} \Pr(N_t \in A) \Pr(T^* = t) \tag{1}$$

where, as the notation suggests, T^* is a random variable with support $\mathbb{N} = \{1, 2, 3, \dots\}$ and, for each fixed $t \in \mathbb{N}$, N_t is a random variable whose support is the state space of Φ. We will explain in Section 2.1 exactly how the distributions of T^* and N_t are related to the Markov transition kernel and the minorization condition.

Representation (1) is appealing from a simulation point of view. Indeed, (1) shows that it is (theoretically) possible to simulate a random variable from π by first drawing from the distribution of T^*, call the result t, and then drawing from the distribution of N_t. This idea is closely related to *perfect* (or *exact*) *sampling*, which is a method of producing independent and identically distributed (iid) samples from correlated (Markovian) simulation. The original perfect sampling algorithms (Propp and Wilson, 1996; Fill, 1998) were designed for large, finite state spaces and, in general, require the user to simulate as many Markov chains as there are points in the state space. While there are some extensions to continuous and unbounded state spaces, these require sophisticated, problem-specific tricks to reduce the number of chains under consideration to a finite number (see e.g., Murdoch and Green, 1998; Hobert, Robert and Titterington, 1999; Mira, Møller and Roberts, 1999; and Fill, Machida, Murdoch and Rosenthal, 2000). A common theme of the perfect samplers that have so far been developed is the use of *coupling*; that is, running chains from different starting points and waiting until they enter the same state at the same time. In contrast, simulating from π via (1) is unrelated to coupling.

The distributions of T^* and N_t are typically not available in closed form. However, as we show in Subsection 2.2, it is *always* possible to make exact draws from the distribution of N_t. Furthermore, when the small set in the minorization condition is the entire state space (and hence Φ is *uniformly ergodic*), T^* is simply a geometric random variable with a known success probability and hence making exact draws from π via (1) is straightforward. Intriguingly, the resulting algorithm is the same as Murdoch and Green's (1998) *multigamma coupler*, which is a perfect sampling algorithm based on *coupling from the past.*

Making exact draws from the distribution of T^* is not straightforward in the case where the small set is not the state space. However, it is possible to estimate the mass function of T^* in this case. More specifically, as we explain in Section 2, $\Pr(T^* = t) \propto \Pr(\tau_\alpha \geq t)$ where τ_α is the first return time to the set α. An unlimited number of iid copies of τ_α can be generated by simply running Φ and counting the number of steps until the chain hits α. These can be used to construct an estimate of the mass function of T^*, which can in turn be used (in conjunction with our ability to simulate N_t) to form an approximation of π that we can actually draw from. Moreover, the (total variation) error in this approximation of π can be assessed using the asymptotic results that we provide. In summary, we show how to construct an arbitrarily precise approximation of π that can be used as the starting distribution for Φ, thereby eliminating the need for burn-in. We illustrate with an example based on the Gibbs sampler.

It is not obvious how to transform the iid copies of τ_α into iid copies of T^*. However, there are alternative methods of drawing from the distribution of T^*, e.g., perfect sampling. We describe a Markov chain whose stationary mass function is that of T^* and that can be simulated using only iid τ_α's. In conjunction with (1) and our ability to draw from the distribution of N_t, a perfect sampling algorithm based on this Markov chain could be used to construct a general algorithm for making exact draws from π.

The rest of the paper is organized as follows. The representation (1) is derived in Section 2. In Section 3, we discuss estimation of the mass function of T^* as well as our application to burn-in. In Section 4, we illustrate our burn-in method using an example involving the Gibbs sampler. The Markov chain Monte Carlo (MCMC) algorithm for simulating T^* is described in Section 5. Finally, in Section 6 we mention a couple of different ways to reuse (or recycle) the simulations that are required to estimate the mass function of T^*.

2 Representing π as a Mixture

2.1 Minorization and the split chain

Let $\mathcal{X} \subseteq \mathbb{R}^p$ for $p \geq 1$ and let $\mathbb{B}(\mathcal{X})$ denote the corresponding Borel σ-algebra. Suppose that $\Phi = \{X_n : n = 0, 1, \dots\}$ is a discrete time, time homogeneous Markov chain with state space $\mathcal{X}$ and Markov transition kernel P; that is, for $x \in \mathcal{X}$ and $A \in \mathbb{B}(\mathcal{X})$

$$P(x, A) = \Pr(X_{n+1} \in A | X_n = x) \; .$$

Let P^n denote the n-step transition kernel based on P; i.e.,

$$P^n(x, A) = \Pr(X_{i+n} \in A | X_i = x) \ .$$

Hence, $P^n(x_0, \cdot)$ is the probability measure of X_n *conditional* on starting the Markov chain at the fixed value x_0. Also, let $P^n(\lambda, \cdot)$ represent the probability measure of X_n given that X_0 is a random variable with probability measure $\lambda(\cdot)$. We assume throughout the paper that Φ is π-irreducible, aperiodic, and positive Harris recurrent where π is the invariant probability measure.

Assume further that a minorization condition has been established for Φ; that is, we have a probability measure $Q(\cdot)$, a small set C for which $\pi(C) > 0$, and an $\varepsilon > 0$ such that for any $x \in C$

$$P(x, A) \geq \varepsilon Q(A) \quad \forall \ A \in \mathbb{B}(\mathcal{X}). \tag{2}$$

This implies that Φ is *strongly aperiodic*. See Meyn and Tweedie (1993, Chapter 5) for a thorough discussion of minorization conditions, including existence issues. Rosenthal (1995) provides a general technique for establishing (2) in the context of MCMC (see also Mykland, Tierney and Yu, 1995).

The random variables T^* and N_t are defined in terms of the *split chain* that is now developed. If (2) holds and $x \in C$ then we can write $P(x, \cdot)$ as a two-component mixture

$$P(x, \cdot) = \varepsilon Q(\cdot) + (1 - \varepsilon) R(x, \cdot) \tag{3}$$

where $R(x, \cdot) = (1-\varepsilon)^{-1}(P(x, \cdot) - \varepsilon Q(\cdot))$ is called the *residual* (or *depleted*) *measure*. When $X_n \in C$, (3) can be used to generate X_{n+1} sequentially as follows. Given $X_n \in C$, generate $\delta_n \sim$ Bernoulli(ε). If $\delta_n = 1$, then draw X_{n+1} from $Q(\cdot)$, else draw X_{n+1} from $R(X_n, \cdot)$.

Athreya and Ney (1978) introduced the *split* chain, $\Phi' = \{(X_n, \delta_n) : n = 0, 1, \dots\}$, which lives on the space $\mathcal{X} \times \{0, 1\}$ and has Markov transition kernel given by

$$P'\left[(x, 0), A \times \delta\right] = \begin{cases} \left[\varepsilon\delta + (1-\varepsilon)(1-\delta)\right] P(x, A) & x \notin C \\ \left[\varepsilon\delta + (1-\varepsilon)(1-\delta)\right] R(x, A) & x \in C \end{cases}$$

and

$$P'\left[(x, 1), A \times \delta\right] = \begin{cases} \left[\varepsilon\delta + (1-\varepsilon)(1-\delta)\right] P(x, A) & x \notin C \\ \left[\varepsilon\delta + (1-\varepsilon)(1-\delta)\right] Q(A) & x \in C \end{cases}$$

where $\delta \in \{0, 1\}$. It is clear that, marginally, the sequence $\{X_n : n = 0, 1, \dots\}$ from the split chain is equivalent (in distribution) to Φ. Also, the measure π' defined by

$$\pi'(A \times \delta) = \pi(A) \left[\varepsilon\delta + (1-\varepsilon)(1-\delta)\right]$$

is invariant for Φ'. The key to our argument is that the set $\alpha := C \times \{1\}$ is an *accessible atom* (Meyn and Tweedie, 1993, p. 100). The (random) times at which Φ' enters α are called *regeneration times* and are the basis of the *regenerative simulation* technique described by Mykland et al. (1995) and Robert (1995).

Before proceeding to the mixture representation, we briefly consider simulation of the split chain. Actually making draws from the residual measure, $R(x, \cdot)$, can be quite difficult. Fortunately, Mykland et al. (1995) provide a simple and clever way of avoiding R altogether. Suppose that $X_n \in C$. Instead of drawing $\delta_n \sim \text{Bernoulli}(\varepsilon)$ and then simulating X_{n+1} conditional on (δ_n, X_n) (which entails simulation from $R(X_n, \cdot)$ if $\delta_n = 0$), one first simulates X_{n+1} conditional on X_n (in the usual way) and then simulates δ_n conditional on (X_n, X_{n+1}). Calculation of $\Pr(\delta_n = 1 | X_n, X_{n+1})$ is straightforward when densities corresponding to P and Q are available.

2.2 Mixture representation

Let τ_α be the first return time to α; that is,

$$\tau_\alpha = \min\{n \geq 1 : (X_n, \delta_n) \in \alpha\} \ .$$

Also, let $\Pr_\alpha(\cdot)$ and $\mathrm{E}_\alpha(\cdot)$ denote probability and expectation conditional on $(X_0, \delta_0) \in \alpha$; that is, conditional on $X_1 \sim Q(\cdot)$. It follows from Kac's Theorem (Meyn and Tweedie, 1993, Theorem 10.2.2) that $\mathrm{E}_\alpha(\tau_\alpha) < \infty$. Hence, we can define a random variable T^* through the tail probabilities of τ_α as follows:

$$\Pr(T^* = t) = \frac{\Pr_\alpha(\tau_\alpha \geq t)}{\mathrm{E}_\alpha(\tau_\alpha)} \tag{4}$$

for $t \in \mathbb{N}$. We are now ready to state the main result.

Theorem 1. *Suppose that Φ is a discrete time, time homogeneous Markov chain with state space $\mathcal{X}$ and Markov transition kernel P. Assume that Φ is π-irreducible, aperiodic, and Harris recurrent with invariant probability measure π. Assume further that the minorization condition (2) holds. Then for any $A \in \mathbb{B}(\mathcal{X})$, we have*

$$\pi(A) = \sum_{t=1}^{\infty} \Pr(N_t \in A) \Pr(T^* = t)$$

where the random variable T^* is defined in (4) and N_t is equal in distribution to X_t given $X_1 \sim Q(\cdot)$ <u>and</u> no regenerations before time t.

Proof. Applying Meyn and Tweedie's (1993) Theorem 10.2.1 to Φ' and using the fact that $\pi'(A \times \{0,1\}) = \pi(A)$, we have

$$\pi(A) = \frac{1}{\mathrm{E}_\alpha(\tau_\alpha)} \sum_{t=1}^{\infty} \Pr_\alpha(X_t \in A, \tau_\alpha \geq t) = \sum_{t=1}^{\infty} \Pr_\alpha(X_t \in A | \tau_\alpha \geq t) \Pr(T^* = t)$$

□

Before working through a simple example, we explain how to simulate from the distribution of N_t. Making draws from the distribution of N_1 is straightforward since $N_1 \sim Q(\cdot)$. Now fix $t > 1$ and suppose that p and q are density functions corresponding to P and Q, respectively. If we start Φ' with $(X_0, \delta_0) \in \alpha$, then the joint density of $(X_1, X_2, \ldots, X_t)$, conditional on no regenerations before time t, is proportional to

$$q(x_1)\{p(x_1, x_2) - \epsilon q(x_2) I_C(x_1)\} \ldots \{p(x_{t-2}, x_{t-1}) - \epsilon q(x_{t-1}) I_C(x_{t-2})\} p(x_{t-1}, x_t) \ , \qquad (5)$$

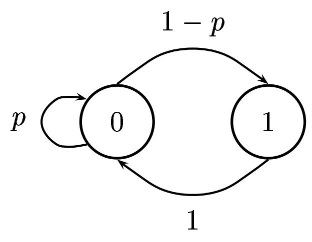
where $I_A(\cdot)$ denotes the indicator function for the set A. Hence, the marginal density of x_t resulting from (5) is exactly the density of N_t. However, simulating directly from the density proportional to (5) is quite challenging.

Fortunately, drawing from the distribution of N_t when $t > 1$ can *always* be accomplished with the following brute force, accept-reject implementation: Take $(X_0, \delta_0) \in \alpha$ and run the split chain for t iterations. This yields $(X_1, \delta_1), \ldots, (X_t, \delta_t)$. If $\delta_1 = \cdots = \delta_{t-1} = 0$, then take X_t as a draw from N_t; otherwise, repeat.

In the next subsection, we consider an artificially simple example in which the distributions of T^* and N_t can be calculated in closed form. In Subsection 2.4, we consider our mixture representation of π in the special case when $C = \mathcal{X}$. The case in which $C \neq \mathcal{X}$ is considered in Subsection 2.5.

2.3 A Simple Example

Suppose Φ evolves according to the following transition diagram:



That is, Φ has state space $\mathcal{X} = \{0, 1\}$ and if we define

$$p_{ij} = P(i, \{j\}) = \Pr(X_{n+1} = j | X_n = i)$$

for $i, j \in \{0, 1\}$ then $p_{00} = p = 1 - p_{01}$ and $p_{10} = 1$. A simple calculation shows that the stationary mass function for this chain is

$$\pi = (\pi_0 \;\; \pi_1)^\top = \begin{pmatrix} \frac{1}{2-p} & \frac{1-p}{2-p} \end{pmatrix}^\top$$

where $\top$ denotes matrix transpose. Assume that $p \in (0, 1/2)$ and let q be the uniform mass function on $\mathcal{X}$; that is, $q(0) = q(1) = 1/2$. It is easy to see that this Markov chain satisfies the following minorization condition

$$p_{0j} \geq \varepsilon\, q(j) \quad \forall\, j \in \mathcal{X}$$

where $\varepsilon = 2p$ and $C = \{0\}$ is the small set. Hence we can split the chain as follows

$$p_{0j} = \varepsilon\, q(j) + (1 - \varepsilon) r(j) \tag{6}$$

where $r(j)$ is a point mass at $j = 1$. Thus given $X_n = 0$, we can generate X_{n+1} sequentially as follows: generate $\delta_n \sim \text{Bernoulli}(\varepsilon)$. If $\delta_n = 1$, then draw X_{n+1} from q; that is, at random from $\{0, 1\}$; else draw X_{n+1} from the mass function r; i.e. set $X_{n+1} = 1$.

Straightforward calculations show that for $t \in \mathbb{N}$

$$\Pr(T^* = t) = \begin{cases} \frac{2p(1-2p)^{\frac{t-1}{2}}}{2-p} & \text{if } t \text{ is odd} \\ \frac{2p(1-p)(1-2p)^{\frac{t-2}{2}}}{2-p} & \text{if } t \text{ is even .} \end{cases}$$

We now derive the distribution of N_t for $t \in \mathbb{N}$. Obviously, N_1 is uniform on $\{0, 1\}$, and the following recursion relation provides a way of calculating $\Pr(N_t = i)$ for $t \in \{2, 3, 4, \dots\}$:

$$\begin{aligned} \Pr(N_t = i) &= \Pr_\alpha(X_t = i | \tau_\alpha \geq t) \\ &= c\Pr_\alpha[X_t = i, (X_{t-1}, \delta_{t-1}) \notin \alpha | \tau_\alpha \geq t - 1] \end{aligned}$$

where $c^{-1} = \Pr_\alpha\left[(X_{t-1}, \delta_{t-1}) \notin \alpha | \tau_\alpha \geq t - 1\right]$. Hence,

$$\begin{aligned} \Pr(N_t = i) &= c \sum_{j=0}^{1} \Pr_\alpha[X_t = i, (X_{t-1}, \delta_{t-1}) \notin \alpha | X_{t-1} = j] \Pr_\alpha(X_{t-1} = j | \tau_\alpha \geq t - 1) \\ &= c \sum_{j=0}^{1} \Pr_\alpha[X_t = i, (X_{t-1}, \delta_{t-1}) \notin \alpha | X_{t-1} = j] \Pr(N_{t-1} = j) \\ &= \begin{cases} c\Pr(N_{t-1} = 1) & \text{if } \; i = 0 \\ c(1 - \varepsilon)\Pr(N_{t-1} = 0) & \text{if } \; i = 1 \end{cases} \end{aligned}$$

We conclude that

$$\Pr(N_t = 0) = \frac{\Pr(N_{t-1} = 1)}{(1-\varepsilon)\Pr(N_{t-1} = 0) + \Pr(N_{t-1} = 1)} = 1 - \Pr(N_t = 1).$$

It follows that

$$\Pr(N_t = 0) = \begin{cases} \frac{1}{2} & \text{if } t \text{ is odd} \\ \frac{1}{2-\varepsilon} & \text{if } t \text{ is even .} \end{cases}$$

Finally,

$$\sum_{t=1}^{\infty} \Pr(N_t = 0)\Pr(T^* = t) = \frac{\Pr(T^* \text{ is odd})}{2} + \frac{\Pr(T^* \text{ is even})}{2-\varepsilon} = \frac{1}{2-p} = \pi_0$$

2.4 The uniformly ergodic case

If Φ satisfies a minorization condition with $C = \mathcal{X}$, then Φ is called *uniformly ergodic*. In this case, $X_n \in C$ for all n and hence $\tau_\alpha \sim \mathcal{G}eo(\varepsilon)$; that is, $\Pr_\alpha(\tau_\alpha = t) = \varepsilon(1-\varepsilon)^{t-1}$ for $t \in \mathbb{N}$. Applying (4) we have that $T^* \sim \mathcal{G}eo(\varepsilon)$ as well. The distribution of N_t is also simple in this case. Indeed, if there were no regenerations before time t, this means that, after having drawn $X_1 \sim Q(\cdot)$, the residual measure was applied $t-1$ consecutive times to get X_t. Hence, simulating from N_t is simply a matter of drawing $X_1 \sim Q(\cdot)$ and then running the Markov chain defined by R for $t-1$ steps. In summary, when $C = \mathcal{X}$, our mixture representation becomes

$$\pi(A) = \sum_{t=1}^{\infty} \varepsilon(1-\varepsilon)^{t-1} R^{t-1}(Q, A) \;, \tag{7}$$

and drawing from π can thus be accomplished using the following algorithm.

Algorithm I:

1. `Simulate` $x_1 \sim Q(\cdot)$ `and` $t \sim \mathcal{G}eo(\varepsilon)$.
2. `If` $t = 1$ `take` x_1, `else run the transition` $x_{n+1} \sim R(x_n, \cdot)$ `for` $n = 1, \cdots, t-1$ `and take` x_t.

Algorithm I is a simple, compact way of describing how to use (1) to sample from π in the uniformly ergodic case. However, if drawing from $R(x, \cdot)$ is very difficult, then Algorithm I is not very practical. Another way to use (1) to sample from π in the uniformly ergodic case is to draw $t \sim \mathcal{G}eo(\varepsilon)$, and then simulate N_t using the brute force method described at the end of Subsection 2.2. Another potential problem with Algorithm I is an ε that is too small. More specifically, since each application of Algorithm I requires running R for about $1/\varepsilon$ iterations, a very small ε could render the algorithm useless from a practical standpoint. In this case, it may be worthwhile to establish an alternative minorization condition in which C is a proper subset of $\mathcal{X}$, and then use the results in Section 3. See Corcoran and Tweedie (2000) for a similar discussion in the context of coupling from the past.

Note that Algorithm I is exactly Murdoch and Green's (1998, p. 486) multigamma coupler, which is based on coupling from the past. By exploiting (3), it is conceptually easy to perform coupling from the past in the uniformly ergodic case. Indeed, if $Q(\cdot)$ is used at time -5, say, then at time -4 all possible chains will have coalesced. Thus, one simply goes back in time until one finds the first use of $Q(\cdot)$ (which will be a geometric number of steps) and then one runs the residual measure forward until time 0. It is interesting that we were able to derive this algorithm with no appeal to coupling (see also Cai, 1999).

While there are some examples of uniformly ergodic MCMC kernels; e.g., the independence Metropolis–Hastings algorithm with fat tail proposals (Mengersen and Tweedie, 1996), these are the exception rather than the rule (Robert and Casella, 1999, Chapter 9). Before moving on to the more challenging case where $C \neq \mathcal{X}$, we briefly describe an alternative way of deriving (7).

Suppose that $K(\cdot,\cdot)$ is a Markov transition kernel corresponding to an irreducible, aperiodic Markov chain with state space $\mathcal{Z}$. Let $\Gamma = \{Z_n : n = 0, 1, \dots\}$ be the discrete time, time homogeneous Markov chain corresponding to the Markov transition kernel given by

$$M(z,\cdot) = \omega\nu(\cdot) + (1-\omega)K(z,\cdot) \tag{8}$$

where $\omega \in (0,1)$ and $\nu(\cdot)$ is a probability measure with support $\mathcal{Z}$. By solving the functional identity $\int \phi(dx)\, M(x,dz)\, dx = \phi(dz)$, Breyer and Roberts (2000) demonstrate that the measure

$$\phi(\cdot) = \sum_{j=1}^{\infty} \omega(1-\omega)^{j-1} K^{j-1}(\nu,\cdot) \tag{9}$$

is an invariant probability measure for Γ. When our original Markov chain Φ satisfies a minorization condition with $C = \mathcal{X}$, (3) holds for all $x \in \mathcal{X}$ and is thus a mixture of the form (8). In that case, (7) follows from (9). (It is interesting to note that Γ is positive recurrent even if the Markov chain associated with K is badly behaved, e.g., transient.)

2.5 The case in which $C \neq \mathcal{X}$

Generally speaking, perfect sampling is much more difficult when $C \neq \mathcal{X}$. There are a few coupling-based perfect samplers that have been developed for this case, such as the catalytic coupling method of Breyer and Roberts (2000), but these are not easy to apply (see also Green and Murdoch, 1999).

We mentioned above that it is always possible to simulate from the distribution of N_t. Not surprisingly, however, drawing from the distribution of T^* is problematic when $C \neq \mathcal{X}$. On the other hand, making iid draws from the distribution of τ_α is straightforward; just take $X_1 \sim Q(\cdot)$, run the split chain, and count how many iterations until the first regeneration. However, despite the simple relationship between their mass functions, it is not clear how to use iid draws from the distribution of τ_α to get iid draws from the distribution of T^*. An alternative strategy for making exact draws from the distribution of T^* is to develop a perfect sampler based on a Markov chain for which T^* is invariant. Such a chain is discussed in Section 5.

In the next section we show how to use our unlimited supply of iid copies of τ_α to construct an *estimate* of the mass function of T^* via (4). This estimate can be used in conjunction with our ability to simulate from N_t to form a precise approximation to π from which we can draw samples. This approximation can be used to address the burn-in issue. These results are illustrated with an example in Section 4.

3 Estimating the Mass Function of T^* via iid Samples

3.1 Monte Carlo estimation

In this section, we consider estimating $p_t = \Pr(T^* = t)$, $t \in \mathbb{N}$, using an iid sample of τ_α's. Note that

$$p_t = \frac{\Pr_\alpha(\tau_\alpha \geq t)}{\mathrm{E}_\alpha(\tau_\alpha)} = \frac{\Pr_\alpha(\tau_\alpha \geq t)}{\sum_{s=1}^{\infty} \Pr_\alpha(\tau_\alpha \geq s)} = \frac{1 - F(t-1)}{1 + \sum_{s=1}^{\infty}[1 - F(s)]}$$

where $F(t) = \Pr_\alpha(\tau_\alpha \leq t)$ is the distribution function of τ_α conditional on $(X_0, \delta_0) \in \alpha$. Let $\tau_{\alpha,1}, \dots, \tau_{\alpha,m}$ denote an iid sample from the distribution of τ_α, and let $F_m(t)$ denote the corresponding empirical distribution function. We consider the following estimate of p_t

$$\hat{p}_t = \frac{1 - F_m(t-1)}{1 + \sum_{s=1}^{\infty}[1 - F_m(s)]} = \frac{1 - F_m(t-1)}{\overline{\tau}_\alpha} \tag{10}$$

where $\overline{\tau}_\alpha$ is the sample mean. Note that $\sum_{t=1}^{\infty} \hat{p}_t = 1$. Hence, $\{\hat{p}_t\}$ will always be a legitimate mass function on $\mathbb{N}$ from which we can sample. In the next subsection, we use asymptotic arguments to get a handle on the error of the estimate $\{\hat{p}_t\}$.

3.2 Approximating the Monte Carlo error

We use the total variation distance between $\{\hat{p}_t\}$ and $\{p_t\}$; i.e., $\sum_{t=1}^{\infty} |\hat{p}_t - p_t|$, as our measure of error. The reason for this will become clear in Subsection 3.3. Let G_1 and G_2 denote two univariate distribution functions. The L_1-*Wasserstein distance* between the probability distributions corresponding to G_1 and G_2 is defined as (Shorack and Wellner, 1986, Chapter 2)

$$d_1(G_1, G_2) = \int_{-\infty}^{\infty} |G_1(x) - G_2(x)|\, dx.$$

The following result shows that $\{\hat{p}_t\}$ is an asymptotically reasonable estimate of the mass function of T^*.

Theorem 2. *For $\{\hat{p}_t\}$ as defined in (10) we have*

$$\sum_{t=1}^{\infty} |\hat{p}_t - p_t| \leq 2 d_1(F_m, F).$$

Hence, $\sum_{t=1}^{\infty} |\hat{p}_t - p_t| \to 0$ a.s. as $m \to \infty$.

Proof. First

$$\begin{aligned}
|\hat{p}_t - p_t| &= \left| \frac{1 - F_m(t-1)}{\overline{\tau}_\alpha} \pm \frac{1 - F(t-1)}{\overline{\tau}_\alpha} - \frac{1 - F(t-1)}{\mathrm{E}_\alpha(\tau_\alpha)} \right| \\
&\leq \frac{|F_m(t-1) - F(t-1)|}{\overline{\tau}_\alpha} + \frac{[1 - F(t-1)]\, |\overline{\tau}_\alpha - \mathrm{E}_\alpha(\tau_\alpha)|}{\overline{\tau}_\alpha \mathrm{E}_\alpha(\tau_\alpha)} \\
&\leq |F_m(t-1) - F(t-1)| + \frac{[1 - F(t-1)]\, |\overline{\tau}_\alpha - \mathrm{E}_\alpha(\tau_\alpha)|}{\mathrm{E}_\alpha(\tau_\alpha)} \\
&\leq |F_m(t-1) - F(t-1)| + \frac{[1 - F(t-1)] \sum_{s=1}^{\infty} |F_m(s) - F(s)|}{\mathrm{E}_\alpha(\tau_\alpha)}
\end{aligned}$$

and hence

$$\sum_{t=1}^{\infty} |\hat{p}_t - p_t| \leq 2 \sum_{t=1}^{\infty} |F_m(t) - F(t)| = 2 \int_{-\infty}^{\infty} |F_m(t) - F(t)|\, dt = 2 d_1(F_m, F). \tag{11}$$

Finally, the fact that $\mathrm{E}_\alpha \tau_\alpha < \infty$ implies that $d_1(F_m, F) \to 0$ a.s. as $m \to \infty$ (Shorack and Wellner, 1986, p. 65). $\square$

Obviously, no matter how large m is, we can never say for certain that $d_1(F_m, F) < \delta$. However, we can use asymptotic results to make statements like $\Pr[d_1(F_m, F) < \delta] \approx 1 - \alpha$. Indeed, Del Barrio, Giné and Matrán (1999) have recently described the first-order asymptotics for the L_1-Wasserstein distance between the empirical and true distribution functions. In particular, their results imply that if $\mathrm{E}_\alpha\left[\tau_\alpha^{2+\varepsilon}\right] < \infty$ for some $\varepsilon > 0$, then

$$\sqrt{m}\, d_1(F_m, F) \xrightarrow{d} \sum_{t=1}^{\infty} |B(F(t))| \tag{12}$$

where $B(s)$, $0 \le s \le 1$, denotes a Brownian bridge process.

Now suppose that $\tau_{\alpha,1}, \dots, \tau_{\alpha,m'}$ is an initial sample of τ_α's with corresponding empirical distribution function $F_{m'}$. Let $u_{m'}$ denote the number of unique values in this sample. Also, let L denote the random variable $\sum_{t=1}^{\infty} |B(F_{m'}(t))|$, which, if m' is large, should have a distribution quite similar to that of $\sum_{t=1}^{\infty} |B(F(t))|$. Simulating the random variable L is quite simple. Indeed, all that is required is $u_{m'}$ values of one realization of standard Brownian motion in $(0,1)$, which can be done sequentially using only univariate normal draws. Hence, it is easy to find c such that

$$\Pr\left[L < c\right] \approx 1 - \alpha.$$

Then if we take $m = 4c^2/\delta^2$, we can say that $\Pr\left[2d_1(F_m, F) < \delta\right] \approx 1 - \alpha$, and hence that $\{\hat{p}_t\}$ is within δ of $\{p_t\}$ with probability approximately equal to $1 - \alpha$.

Before describing how the results of Subsections 3.1 and 3.2 can be used to avoid burn-in, we briefly discuss the three forms of approximation used in the above argument. The three approximations are: (i) the asymptotic approximation in (12); (ii) the use of $\sum_{t=1}^{\infty} |B(F_{m'}(t))|$ in place of $\sum_{t=1}^{\infty} |B(F(t))|$; and (iii) the estimation of the quantile of L. The most bothersome of these is certainly (ii). It may be possible to dispense with approximations (ii) and (iii), and simplify the method at the same time. Specifically, let $\mathcal{G}$ denote the class of distribution functions corresponding to discrete random variables with support $\mathbb{N}$ that have a finite $2 + \varepsilon$ moment. Consider the random variables

$$L_G = \sum_{t=1}^{\infty} |B(G(t))|$$

as G ranges over $\mathcal{G}$. Suppose there exists a G^* such that L_{G^*} is stochastically larger then any other L_G. Then we could simply use L_{G^*} in place of $\sum_{t=1}^{\infty} |B(F(t))|$. This would remove approximation (ii) and would alleviate the need for the initial sample of m' τ_α's. Furthermore, the quantiles of L_{G^*} could be tabulated and this would obviate approximation (iii). All of this suggests that an investigation into the possible existence of a "dominating" G^* could be well worth the effort. We do not explore this possibility any further in this paper.

3.3 An application to burn-in

We have assumed that Φ is π-irreducible, strongly aperiodic, and Harris recurrent where π is the invariant probability measure. It follows that for every initial probability measure $\lambda(\cdot)$ we have

$$\|P^n(\lambda,\cdot) - \pi(\cdot)\| \to 0 \quad \text{as} \quad n \to \infty, \tag{13}$$

where $\|\cdot\|$ represents total variation distance (Meyn and Tweedie, 1993, Section 13.3). Moreover, $\|P^n(\lambda,\cdot) - \pi(\cdot)\|$ is non-increasing in n. (Note that in most applications of MCMC, $\lambda(\cdot)$ is degenerate at a point x_0.)

An important problem in the implementation and comparison of MCMC algorithms is *burn-in* (time), which can be described formally as follows. Given λ, we want to find a value n^* such that

$$\|P^{n^*}(\lambda,\cdot) - \pi(\cdot)\| < \delta \tag{14}$$

for some arbitrarily small, prespecified $\delta > 0$. If (14) holds, then the marginal distribution of X_n (conditional on $X_0 \sim \lambda$) is within δ of π for all $n \geq n^*$. Therefore, n^* is a reasonable time to start sampling the Markov chain. We are, of course, assuming that perfect sampling from Φ is not an viable option. Otherwise, "getting close" to π would be a mute point.

It is possible to find n^* by constructing exact upper bounds on $\|P^n(\lambda,\cdot) - \pi(\cdot)\|$ through drift and minorization conditions (Rosenthal, 1995; Roberts and Tweedie, 1999). The phrase "difficult theoretical analysis" is used by Fill, Machida, Murdoch and Rosenthal (2000) to describe this method. While this approach has been successfully applied to a few relatively simple MCMC algorithms (Roberts and Rosenthal, 1999; Jones and Hobert, 2000), it is not clear how useful it will be in more complex situations. We now present a slightly less rigorous and much simpler alternative to calculating n^* that is based on our mixture representation of π.

Let $\{p_t\}$ and $\{\hat{p}_t\}$ be as in Subsection 3.1. Also, let $N_t(\cdot)$ denote the probability measure corresponding to N_t. Consider using

$$\hat{\pi}(\cdot) = \sum_{t=1}^{\infty} N_t(\cdot)\hat{p}_t$$

as the initial distribution. Starting Φ with $X_0 \sim \hat{\pi}(\cdot)$ is straightforward from a practical standpoint. Furthermore,

$$\|\hat{\pi}(\cdot) - \pi(\cdot)\| = \left\| \sum_{t=1}^{\infty} N_t(\cdot)\hat{p}_t - \pi(\cdot) \right\| = \left\| \sum_{t=1}^{\infty} N_t(\cdot)\hat{p}_t - \sum_{t=1}^{\infty} N_t(\cdot)p_t \right\| \leq \sum_{t=1}^{\infty} |\hat{p}_t - p_t|.$$

Therefore, if $\sum_{t=1}^{\infty} |\hat{p}_t - p_t| < \delta$, then our initial probability measure is within δ of π and *burn-in is unnecessary*. Unfortunately, we can never say with certainty that $\sum_{t=0}^{\infty} |\hat{p}_t - p_t| < \delta$. However, we can use the results in Subsection 3.2 to say that this is true with high probability (approximately). In summary, suppose that the methods of Subsection 3.2 are used to construct $\{\hat{p}_t\}$ such that

$$\Pr\left[\sum_{t=1}^{\infty} |\hat{p}_t - p_t| < \delta\right] \approx 1 - \alpha$$

where α and δ are small. Then if the Markov chain Φ is started with $X_0 \sim \hat{\pi}(\cdot)$, one can begin sampling the Markov chain right away.

Obviously, it would be nice to be able to say that $\|\hat{\pi}(\cdot) - \pi(\cdot)\| < \delta$ for certain. On the other hand, our method requires far less "difficult theoretical analysis" than the exact methods of Rosenthal (1995) and Roberts and Tweedie (1999). In the next section, we illustrate the use of our method with a realistic example.

4 An Example

4.1 Setting

The model we consider is a Bayesian hierarchical version of the usual normal theory one-way random effects model. First, conditional on $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)^\top$ and λ_e, the data, Y_{ij}, are assumed independent with

$$Y_{ij}|\theta_i, \lambda_e \sim \mathrm{N}(\theta_i, \lambda_e^{-1})$$

where $i = 1, \dots, K$ and $j = 1, \dots, w_i$. At the second stage, conditional on μ and λ_θ, $\boldsymbol{\theta}$ and λ_e are independent with

$$\boldsymbol{\theta}|\mu, \lambda_\theta \sim \mathrm{N}(\mu\mathbf{1}, \lambda_\theta^{-1}\mathbf{I}) \quad \text{and} \quad \lambda_e \sim \mathrm{Exp}(1),$$

where $\mathbf{1}$ is a $K \times 1$ column vector of ones and $\mathbf{I}$ is a $K \times K$ identity matrix. Finally, at the third stage, μ and λ_θ are assumed independent with

$$\mu \sim \mathrm{N}(0,1) \quad \text{and} \quad \lambda_\theta \sim \mathrm{Exp}(1).$$

We assume that $K \geq 2$ and that $w_i \geq 2$ for all i. The posterior density is characterized by

$$\pi(\boldsymbol{\theta}, \mu, \lambda_e, \lambda_\theta|y) \propto f(y|\boldsymbol{\theta}, \lambda_e) f(\boldsymbol{\theta}|\mu, \lambda_\theta) f(\lambda_e) f(\mu) f(\lambda_\theta) \tag{15}$$

where y is a vector containing all of the observed data, and f denotes a generic density. The integrals required for inferences through this posterior are not available in closed form. Thus, we must resort to (possibly) high dimensional numerical integration, analytical approximations (e.g. Laplace approximation), or MCMC techniques like the Gibbs sampler.

4.2 A two-stage Gibbs sampler

The goal is to make draws from the posterior density $\pi(\boldsymbol{\theta}, \mu, \lambda_e, \lambda_\theta | y)$. The data, y, is fixed at the observed (or simulated in our case) value. We employ a two-component (block) Gibbs sampler that has been analyzed extensively in Hobert and Geyer (1998) and Jones and Hobert (2000). More specifically, we consider the Gibbs sampler that updates $\boldsymbol{\lambda} = (\lambda_\theta, \lambda_e)^\top$ and then $\boldsymbol{\xi} = (\theta_1, \dots, \theta_K, \mu)^\top$; that is, the update scheme looks like

$$(\boldsymbol{\lambda}', \boldsymbol{\xi}') \to (\boldsymbol{\lambda}, \boldsymbol{\xi}') \to (\boldsymbol{\lambda}, \boldsymbol{\xi})$$

where we are using $(\boldsymbol{\lambda}', \boldsymbol{\xi}')$ and $(\boldsymbol{\lambda}, \boldsymbol{\xi})$ to denote the current and future states, respectively, in order to avoid excessive subscripting. It follows from the form of (15) that λ_e and λ_θ are conditionally independent given $\boldsymbol{\xi}$. Hence, the Markov transition density for our Gibbs sampler is

$$k(\boldsymbol{\lambda}, \boldsymbol{\xi} | \boldsymbol{\lambda}', \boldsymbol{\xi}') = \pi(\lambda_\theta | \boldsymbol{\xi}', y) \; \pi(\lambda_e | \boldsymbol{\xi}', y) \; \pi(\boldsymbol{\xi} | \boldsymbol{\lambda}, y) \tag{16}$$

where $\pi(\lambda_\theta | \boldsymbol{\xi}', y)$ and $\pi(\lambda_e | \boldsymbol{\xi}', y)$ are gamma densities and $\pi(\boldsymbol{\xi} | \boldsymbol{\lambda}, y)$ is multivariate normal. See Hobert and Geyer (1998) for the details.

4.3 The minorization condition

Let $\overline{y}_i = w_i^{-1} \sum_{j=1}^{w_i} y_{ij}$, $\mathrm{SSE} = \sum_{i,j} (y_{ij} - \overline{y}_i)^2$, $w = \sum_i w_i$ and let c_1 and c_2 denote two fixed positive numbers. Also, define

$$v_1(\boldsymbol{\xi}) = \sum_{i=1}^{K} (\theta_i - \mu)^2 \quad \text{and} \quad v_2(\boldsymbol{\xi}) = \sum_{i=1}^{K} w_i (\theta_i - \overline{y}_i)^2 \; .$$

Jones and Hobert (2000) established that

$$\pi(\lambda_\theta | \boldsymbol{\xi}', y) \pi(\lambda_e | \boldsymbol{\xi}', y) \pi(\boldsymbol{\xi} | \boldsymbol{\lambda}, y) \geq \varepsilon q(\boldsymbol{\lambda}, \boldsymbol{\xi}) \tag{17}$$

for all $(\boldsymbol{\lambda}', \boldsymbol{\xi}') \in C$ where

$$C = \{(\boldsymbol{\lambda}, \boldsymbol{\xi}) : v_1(\boldsymbol{\xi}) < c_1 \;\text{ and }\; v_2(\boldsymbol{\xi}) < c_2\}$$

and the function $q(\boldsymbol{\lambda}, \boldsymbol{\xi})$ is a density on $\mathbb{R}_+^2 \times \mathbb{R}^{K+1}$ whose precise form is described below. Clearly, integration of (17) leads directly to a minorization condition of the form (2). We now describe the density $q(\boldsymbol{\lambda}, \boldsymbol{\xi})$.

Let Gamma$(\alpha, \beta; r)$ denote the value of the Gamma(α, β) density at the point $r > 0$; i.e., Gamma$(\alpha, \beta; r) = \frac{\beta^\alpha}{\Gamma(\alpha)} r^{\alpha-1} e^{-r\beta}$. Define two functions as follows:

$$g(\lambda_\theta) = \begin{cases} \text{Gamma}\left(\frac{K}{2}+1, 1; \lambda_\theta\right) & 0 < \lambda_\theta < \lambda_\theta^* \\ \text{Gamma}\left(\frac{K}{2}+1, \frac{c_1}{2}+1; \lambda_\theta\right) & \lambda_\theta \geq \lambda_\theta^* \end{cases}$$

and

$$h(\lambda_e) = \begin{cases} \text{Gamma}\left(\frac{w}{2}+1, \frac{\text{SSE}}{2}+1; \lambda_e\right) & 0 < \lambda_e < \lambda_e^* \\ \text{Gamma}\left(\frac{w}{2}+1, \frac{\text{SSE}+c_2}{2}+1; \lambda_e\right) & \lambda_e \geq \lambda_e^* \end{cases}$$

where

$$\lambda_\theta^* = (K+2)\log\left(1+\frac{c_1}{2}\right)/c_1 \quad \text{and} \quad \lambda_e^* = (w+2)\log\left(1+\frac{c_2}{2+\text{SSE}}\right)/c_2 \; .$$

If we set $\varepsilon_1 = \int_0^\infty g(\lambda_\theta)\, d\lambda_\theta$ and $\varepsilon_2 = \int_0^\infty h(\lambda_e)\, d\lambda_e$, then $g(\lambda_\theta)/\varepsilon_1$ is a density in λ_θ, and likewise, $h(\lambda_e)/\varepsilon_2$ is a density in λ_e. Moreover, simulating from these densities is quite straightforward as each can be written as a mixture of two truncated Gamma densities.

The density $q(\boldsymbol{\lambda}, \boldsymbol{\xi})$ is defined by

$$q(\boldsymbol{\lambda}, \boldsymbol{\xi}) = \left[\frac{g(\lambda_\theta)}{\epsilon_1}\right]\left[\frac{h(\lambda_e)}{\epsilon_2}\right]\pi(\boldsymbol{\xi}|\boldsymbol{\lambda}, y)$$

and ε from (17) is defined as $\varepsilon = \varepsilon_1\varepsilon_2$.

4.4 Constructing $\hat{\pi}(\cdot)$

We now illustrate our method using the data in Table 1, which were simulated according to the hierarchical model described in Subsection 4.1. Our goal will be to figure out how large a sample of τ_α's is required in order to say that

$$\|\hat{\pi}(\cdot) - \pi(\cdot)\| = \left\|\sum_{t=1}^{\infty} N_t(\cdot)\hat{p}_t - \pi(\cdot)\right\| < 0.01 \tag{18}$$

with (approximate) probability 0.95. Thus, in the notation of Subsection 3.2 we are using $\delta = 0.01$ and $\alpha = 0.05$. (Note that we are abusing notation a bit here in letting π represent both the posterior density and the corresponding probability measure.)

For this particular data set, we use $c_1 = 2$ and $c_2 = 3$ so that the small set is defined as

$$C = \{(\boldsymbol{\lambda}, \boldsymbol{\xi}) : v_1(\boldsymbol{\xi}) < 2 \;\text{ and }\; v_2(\boldsymbol{\xi}) < 3\} \; .$$

Recall that

$$\tau_\alpha = \min\left\{n \geq 1 : \Phi'_n \in \alpha\right\}$$

when $(\boldsymbol{\lambda}_1, \boldsymbol{\xi}_1) \sim q(\boldsymbol{\lambda}, \boldsymbol{\xi})$. First, an initial sample of $m' = 100,000$ iid τ_α's was drawn. The largest value was 45 and only about 1% were larger than 16. Let $F_{m'}$ denote the empirical distribution function based on this initial sample. We then simulated 100,000 values of $L = \sum_{t=1}^{\infty} |B(F_{m'}(t))|$. The 5,000th largest value in this sample was 7.211, and we concluded from this that $\Pr[L < 7.211] \approx 0.95$. Consequently, if we take a sample of size

$$m = \frac{4c^2}{\delta^2} = \frac{4(7.211)^2}{(0.01)^2} \approx 2.08 \times 10^6$$

τ_α's, then we can assert that (18) is true with (approximate) probability 0.95.

We then simulated an additional $m = 2.08 \times 10^6$ iid τ_α's. (This took about 30 minutes using a Sun Ultra 5 with 256 MB RAM and a program written in the Ox language.) Table 2 describes this sample. (In order to check our original estimate of c, we simulated 100,000 copies of the L corresponding to this new estimate of F, and the 5,000th largest value in this sample was 7.210.) Table 3 contains the corresponding estimate of the mass function of the random variable T^*.

Choosing $X_0 \sim \hat{\pi}(\cdot) = N_t(\cdot)\hat{p}_t$ can be done as follows. Draw a random variable from the mass function in Table 3, call the result t. If $t = 1$, take $X_0 \sim q(\boldsymbol{\lambda}, \boldsymbol{\xi})$. If $t > 1$, get $(\boldsymbol{\lambda}_1, \boldsymbol{\xi}_1) \sim q(\boldsymbol{\lambda}, \boldsymbol{\xi})$ and run the (split) Gibbs chain for an additional $t-1$ steps. If $\delta_1 = \cdots = \delta_{t-1} = 0$, set $X_0 = (\boldsymbol{\lambda}_t, \boldsymbol{\xi}_t)$. If not, repeat the previous step until $\delta_1 = \cdots = \delta_{t-1} = 0$. That is, get a new (independent) $(\boldsymbol{\lambda}_1, \boldsymbol{\xi}_1) \sim q(\boldsymbol{\lambda}, \boldsymbol{\xi})$ and so on.

5 MCMC simulation of T^*

Let Z and Z^* be univariate random variables with support $\mathbb{N}$. Suppose that Z has a finite mean and that $\Pr(Z^* = t) \propto \Pr(Z \geq t)$. In this section, we develop a generic Markov chain whose stationary mass function is that of Z^* and which can be simulated using only iid copies of Z.

5.1 A Markov chain from the tails

For $i \in \mathbb{N}$, let $Z^{(i)}$ denote the i-truncated version of Z; that is,

$$\Pr(Z^{(i)} = j) = \frac{\Pr(Z = j)}{\Pr(Z \geq i)} \qquad j = i, i+1, i+2, \ldots$$

Thus, $Z^{(1)} \stackrel{d}{=} Z$. Consider a discrete time, time homogeneous Markov chain $\Psi = \{Y_n : n = 0, 1, \dots\}$, with state space $\mathbb{N}$, that evolves as follows. If the current value is $Y_n = i$, then $Y_{n+1} = Z^{(i)} - (i-1)$; that is, we generate a random variable from the i-truncated version of Z, and then subtract $i-1$ so that the support remains equal to $\mathbb{N}$ at every step. Here is a general result regarding Ψ.

Theorem 3. *The Markov chain Ψ is irreducible and aperiodic. Furthermore, the sequence $\pi_i = \Pr(Z \geq i)$, $i \in \mathbb{N}$, is invariant and hence Ψ is positive recurrent if and only if $E(Z) < \infty$. Lastly, Ψ is reversible with respect to π.*

Proof. Irreducibility and aperiodicity are obvious. The formal transition probabilities are given by

$$\Pr(Y_{n+1} = j | Y_n = i) = p_{ij} = \frac{\Pr(Z = i + j - 1)}{\Pr(Z \geq i)}$$

for $i, j \in \mathbb{N}$. Now,

$$\sum_{i=1}^{\infty} \pi_i \, p_{ij} = \sum_{i=1}^{\infty} \Pr(Z = i + j - 1) = \Pr(Z \geq j) = \pi_j$$

and hence the sequence $\{\pi_i\}$ is invariant. If this invariant sequence has finite total mass; i.e., if $\sum_{i=1}^{\infty} \pi_i < \infty$, then the normalized version of π_i is the unique invariant mass function and the Markov chain is positive recurrent. If, on the other hand, the sum diverges, then the chain is either null recurrent or transient. Since $\sum_{i=1}^{\infty} \pi_i = E(Z)$, the chain is positive recurrent if and only if Z has finite expectation. Finally, note that

$$\pi_i \, p_{ij} = \Pr(Z = i + j - 1) = \pi_j \, p_{ji}$$

and this establishes the reversibility. $\square$

EXAMPLE 5.1. Suppose that $Z \sim \mathcal{G}eo(p)$; that is, for $i \in \mathbb{N}$, $\Pr(Z = i) = p(1-p)^{i-1}$. Because of the memoryless property of the geometric distribution, $Z^{(i)} - i + 1$ has the same distribution as Z, and thus Ψ is simply an iid sequence of $\mathcal{G}eo(p)$ random variables.

$\|$

5.2 Application

We now return to the problem of drawing from the distribution of T^*. If we construct the chain Ψ using the random variable τ_α of Section 2 in place of Z, then the mass function of T^* is invariant. Note that it is actually possible to simulate Ψ despite the fact that the mass function of τ_α is not available in closed form. Specifically, if the current value of the chain is $Y_n = i$, then in order to simulate Y_{n+1}, we need to make a draw from the i-truncated version of τ_α. A brute force way to do this is to simulate τ_α's until we get one $\geq i$.

This is an MCMC algorithm and hence does not produce exact draws from T^*. However, if one could develop a perfect sampler based on Ψ, then the resulting exact draws from T^* could be used in conjunction with (1) to make exact draws from π. The two major stumbling blocks are (i) T^* is not bounded above, and (ii) the transition kernel for Ψ is not available in closed from.

6 A Recycling Technique

In Subsection 3.1, we discussed a method of estimating the mass function of T^* using iid draws from the distribution of τ_α. A byproduct of this method are sequences of the form $x_0^j, \ldots, x_{\tau_{\alpha,j}}^j$ for $j = 1, \ldots, m$. Each of these sequences was the result of simulating the split chain (started with $(X_0, \delta_0) \in \alpha$) until the first regeneration occured. Therefore, these sequences can be recycled when draws from N_t are required. Specifically, for fixed $t > 1$, any x_t^j taken from a sequence such that $\tau_{\alpha,j} \geq t$ is a realization of N_t.

Another way to recycle the sequences is as follows. Suppose we wish to estimate $\mathrm{E}^\pi\left[h(X)\right]$. Now consider the decomposition

$$\begin{aligned} \mathrm{E}^\pi\left[h(X)\right] &= \sum_{t=1}^{\infty} P(T^* = t)\, \mathrm{E}\left[h(N_t)\right] \\ &= \frac{\sum_{t=1}^{\infty} \mathrm{E}_\alpha\left[h(X_t)\, I(\tau_\alpha \geq t)\right]}{\mathrm{E}_\alpha[\tau_\alpha]} . \end{aligned}$$

The denominator can be estimated with $\overline{\tau}_\alpha$. The expectation in the numerator can be estimated with

$$\frac{1}{m} \sum_{j=1}^{m} h(x_t^j)\, I(\tau_{\alpha,j} \geq t) .$$

Note that if $\tau_{\alpha,j} < t$, then x_t^j was not simulated. We simply define $h(x_t^j)$ to be zero whenever $\tau_{\alpha,j} < t$. Putting all of this together gives the following approximation

$$\mathrm{E}^\pi\left[h(X)\right] \approx \frac{\sum_{j=1}^{m} \sum_{t=1}^{\tau_{\alpha,j}} h(x_t^j)}{m \overline{\tau}_\alpha} .$$

Acknowledgments.

The authors are grateful to Eric Moulines for pointing out the connection between the mixture representation and Kac's Theorem, and also to Richard Tweedie for a discussion regarding mixtures of Markov kernels in August 1999. Discussions with Laird Breyer, Gareth Roberts and Richard Tweedie during the Ambleside meeting in September 2000 are also gratefully acknowledged!

References

Athreya, K. B. and Ney, P. (1978). A new approach to the limit theory of recurrent Markov chains, *Transactions of the American Mathematical Society* **245**: 493–501.

Breyer, L. A. and Roberts, G. O. (2000). Some multistep coupling constructions for Markov chains, *Technical report*, University of Lancaster, UK.

Cai, H. (1999). Exact sampling using auxiliary variables, *Technical report*, University of Missouri, St. Louis.

Corcoran, J. N. and Tweedie, R. L. (2000). Perfect sampling of Harris recurrent Markov chains, *Annals of Applied Probability,* to appear.

Del Barrio, E., Giné, E. and Matrán, C. (1999). Central limit theorems for the Wasserstein distance between the empirical and the true distributions, *Annals of Probability* **27**: 1009–1071.

Fill, J. A. (1998). An interruptible algorithm for perfect sampling via Markov chains, *Annals of Applied Probability* **8**: 131–162.

Fill, J. A., Machida, M., Murdoch, D. J. and Rosenthal, J. S. (2000). Extension of Fill's perfect rejection sampling algorithm to general chains, *Technical report*, The Johns Hopkins University.

Green, P. J. and Murdoch, D. (1999). Exact sampling for Bayesian inference: Towards general purpose algorithms, *in* J. O. Berger, J. M. Bernardo, A. P. Dawid, D. V. Lindley and A. F. M. Smith (eds), *Bayesian Statistics 6*, Oxford University Press, Oxford, pp. 302–321.

Hobert, J. P. and Geyer, C. J. (1998). Geometric ergodicity of Gibbs and block Gibbs samplers for a hierarchical random effects model, *Journal of Multivariate Analysis* **67**: 414–430.

Hobert, J. P., Robert, C. P. and Titterington, D. M. (1999). On perfect simulation for some mixtures of distributions, *Statistics and Computing* **9**: 287–298.

Jones, G. L. and Hobert, J. P. (2000). Upper bounds on the distance to stationarity for the block Gibbs sampler for a hierarchical random effects model, *Technical report*, University of Florida.

Mengersen, K. L. and Tweedie, R. L. (1996). Rates of convergence of the Hastings and Metropolis algorithms, *The Annals of Statistics* **24**: 101–121.

Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability*, Springer-Verlag, London.

Mira, A., Møller, J. and Roberts, G. (1999). Perfect slice samplers, *Technical report*, Universitá dell'Insubria, Varese, Italy.

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

Mykland, P., Tierney, L. and Yu, B. (1995). Regeneration in Markov chain samplers, *Journal of the American Statistical Association* **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.

Robert, C. P. (1995). Convergence control methods for Markov chain Monte Carlo algorithms, *Statistical Science* **10**: 231–253.

Robert, C. P. and Casella, G. (1999). *Monte Carlo Statistical Methods*, Springer, New York.

Roberts, G. O. and Rosenthal, J. S. (1999). Convergence of slice sampler Markov chains, *Journal of the Royal Statistical Society,* Series B **61**: 643–660.

Roberts, G. O. and Tweedie, R. L. (1999). Bounds on regeneration times and convergence rates for Markov chains, *Stochastic Processes and their Applications* **80**: 211–229.

Rosenthal, J. S. (1995). Minorization conditions and convergence rates for Markov chain Monte Carlo, *Journal of the American Statistical Association* **90**: 558–566.

Shorack, G. R. and Wellner, J. A. (1986). *Empirical Processes with Applications to Statistics*, John Wiley and Sons, New York.

Table 1: Simulated Data

Group	1	2	3	4	5	6
w_i	5	5	5	10	10	10
$\bar{y}_i$	-0.10780	-0.16943	0.00552	-0.22936	-0.25921	-0.14966

$$w = \sum_{i=1}^{6} w_i = 45$$

$$\text{SSE} = \sum_{i=1}^{6} \sum_{j=1}^{w_i} (y_{ij} - \bar{y}_i)^2 = 14.928$$

Table 2: Frequencies of outcomes in a sample of 2.08×10^6 iid τ_α's

τ_α	freq.	τ_α	freq.	τ_α	freq.	τ_α	freq.	τ_α	freq.
1	537891	11	28162	21	1581	31	82	41	3
2	390913	12	21219	22	1142	32	81	42	1
3	290861	13	15848	23	829	33	47	43	5
4	216527	14	11673	24	656	34	33	44	3
5	162093	15	8823	25	482	35	26	45	3
6	121696	16	6565	26	341	36	28	46	1
7	91230	17	5023	27	267	37	22	47	2
8	68136	18	3795	28	198	38	14	48	0
9	50847	19	2761	29	161	39	5	49	0
10	37686	20	2109	30	124	40	4	50	1

Table 3: Estimate of the mass function of T^* (to seven decimal places)

t	$\hat{p}_t$	t	$\hat{p}_t$
1	.2541461	26	.0001774
2	.1884236	27	.0001357
3	.1406596	28	.0001031
4	.1051206	29	.0000789
5	.0786641	30	.0000593
6	.0588587	31	.0000441
7	.0439892	32	.0000341
8	.0328422	33	.0000242
9	.0245169	34	.0000185
10	.0183041	35	.0000144
11	.0136995	36	.0000112
12	.0102585	37	.0000078
13	.0076658	38	.0000051
14	.0057294	39	.0000034
15	.0043031	40	.0000028
16	.0032251	41	.0000023
17	.0024229	42	.0000020
18	.0018092	43	.0000018
19	.0013455	44	.0000012
20	.0010082	45	.0000009
21	.0007505	46	.0000005
22	.0005573	47	.0000004
23	.0004178	48	.0000001
24	.0003165	49	.0000001
25	.0002363	50	.0000001